The pan-NLR'ome of *Arabidopsis thaliana*

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Anna-Lena Van de Weyer
aus Bad Kissingen

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.03.2019
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Thorsten Nürnberger

Acknowledgements

I would like to thank my thesis advisors Prof. Detlef Weigel and Prof. Thorsten Nürnberger for not only their continuous support during my PhD studies but also for their critical questions and helpful ideas that shaped and developed my project. Furthermore, I would like to thank Dr. Felicity Jones for being part of my Thesis Advisory Committee, for fruitful discussions and for her elaborate questions.

I would like to express my sincere gratitude to Dr. Felix Bemm, who worked closely with me during the last years, who continuously challenged and promoted me and without whom I would not be where and who I am now. In the mindmap for my acknowledgement I wrote down 'Felix: thanks for everything'. That says it all.

Further thanks goes to my collaborators Dr. Freddy Monteiro and Dr. Oliver Furzer for their essential contributions to experiments and analyses during all stages of my PhD. Their sheer endless knowledge of NLR biology amplified my curiosity and enthusiasm to learn. Freddy deserves a special thank you for his awesome collaborative spirit and for answering all my NLR related newbie questions.

I am grateful for my colleague and friend Clemens Weiß, who not only helped me improve my knowledge of bioinformatics, but especially provided vital emotional support, esprit and fun at all times. Thanks to my office mates Max Collenberg, Christian Kubica, Dr. Leily Rabbani, and Dr. Ilja Bezrukov for fruitful discussions, continuous chocolate supply, and the fun that we had together. Special thanks goes to Dr. Dagmar Sigurdadottir and Dr. Rebecca Schwab for their endless support with everything that needed to be organized.

Last, but certainly not least, I want to thank my family. Thank you so much mum Claudia and dad Klaus for always supporting and believing in me. I am so grateful for my husband Nico, for his unconditional support. Thanks for being a stay-at-home dad, allowing me to pursue my research, thanks for the little gifts (how would I have survived without my daily 'Sesamstange'), thanks for listening to hours of bioinformatics gibberish and for your endless love. Special thanks goes to my son Maximilian who was the main driving force for me to finish my PhD as quickly as possible, whose smile always makes the sun shine and whose motivating words will never be forgotten (Mamamamamam, pfffffrrt,bababa!)

Summary

Plants are the major nutritional component of the human diet, provide us with shelter, fuel, and enjoyment. Substantial yield loss is caused by plant diseases transmitted by bacteria, fungi, and oomycete pathogens. Plants have an elaborate innate immune system to fight threatening pathogens, relying to a great extend on highly variable resistance (R) genes. R genes often encode intracellular nucleotide-binding leucine-rich repeat receptors (NLRs) that directly or indirectly recognize pathogens by the presence or the activity of effector proteins in the plants' cells. NLRs contain variable N-terminal domains, a central nucleotide-binding (NB) domain, and C-terminal leucine-rich repeats (LRRs). The N-terminal domains can be used to distinguish between the evolutionary conserved NLR classes TNL (with a toll/interleucin-1 receptor homology (TIR) domain), CNL (with a coiled-coil (CC) domain), and RNL (with an RPW8 domain). The architectural diversity is increased by additional integrated domains (IDs) found in different positions. Plant species have between a few dozen and several hundred NLRs. The intraspecific R gene diversity is also high, and the still few known NLRs responsible for long-term resistance are often accession-specific. Intraspecific NLR studies to date suffer from several shortcomings: The pan-NLR'omes (the collection of all NLR genes and alleles occurring in a species) can often not be comprehensively described because too few accessions are analyzed, and NLR detection is essentially always guided by reference genomes, which biases the detection of novel genes and alleles. In addition, inappropriate or immature bioinformatics analysis pipelines may miss NLRs during the assembly or annotation phase, or result in erroneous NLR annotations. Knowing the pan-NLR'ome of a plant species is key to obtain novel resistant plants in the future. I created an extensive and reliable database that defines the near-complete pan-NLR'ome of the model plant *Arabidopsis thaliana*. Efforts were focused on a panel of 65 diverse accessions and applied state-of-the-art targeted long read sequencing (SMRT RenSeq). My analysis pipeline was designed to include optimized methods that could be applied to any SMRT RenSeq project. In the first part of my thesis I set quality control standards for the assembly of NLR-coding genomic fragments. I further introduce a novel and thorough gene annotation pipeline, supported by careful manual curation. In the second part, I present the manuscript reporting the saturated near-complete *A. thaliana* pan-NLR'ome. The species-wide high NLR diversity is revealed on the domain architecture level, and the usage of novel IDs is highlighted. The core NLR complement is defined and presence-absence polymorphisms in non-core NLRs are described. Furthermore, haplotype saturation is shown, selective forces are quantified, and evolutionary coupled co-evolving NLRs are detected. The method optimization results show that final NLR assembly quality is mainly influenced by the amount and the quality of input sequencing data. The results further show that manual curation of automated NLR predictions are crucial to prevent frequently occurring misannotations. The saturation of an NLR'ome has not been shown in any plant species so far, thus this study provides an unprecedented view on intraspecific NLR variation, the core NLR complement, and the evolutionary trajectories of NLRs. IDs are more frequently used than known before, suggesting a pivotal role of noncanonical NLRs in plant-pathogen interactions. This work sets new

standards for the analysis of gene families at the species level. Future NLR'ome projects applied to important crop species will profit from my results and the easy-to-adopt analysis pipeline. Ultimately, this will extend our knowledge of intraspecific NLR diversity beyond few reference species or genomes, and will facilitate the detection of functional NLRs, to be used in disease resistance breeding programs.

Zusammenfassung

Pflanzen sind Hauptbestandteil der menschlichen Ernährung, liefern Schutz, Kraftstoffe, und Erholung. Durch Pflanzenkrankheiten die von Bakterien, Pilzen und Oomyceten übertragen werden können, werden beträchtliche Ertragseinbußen verursacht. Pflanzen haben ein ausgefeiltes angeborenes Immunsystem um gefährliche Pathogene zu bekämpfen. Dabei sind sie hauptsächlich auf die hoch-variablen Resistenz (R) Gene angewiesen. R Gene kodieren oft intrazelluläre nukleotid-bindende Rezeptoren mit leucin-reichen repetitiven Regionen (nucleotide-binding leucine-rich repeat receptors, NLRs), die Pathogene direkt oder indirekt wahrnehmen können indem sie die Präsenz oder Aktivität von Effektor Proteinen in den pflanzlichen Zellen detektieren. NLRs beinhalten variable N-terminale Domänen, eine zentrale Nukleotid-bindende (nucleotide-binding, NB) Domäne, und C-terminale Leucin-reiche repetitiven Regionen (leucine-rich repeats, LRRs). Die N-terminalen Domänen können verwendet werden um zwischen den evolutionär konservierten NLR Klassen TNL (mit einer toll/interleucin-1 Rezeptor homologen (TIR) Domäne), CNL (mit einer superspiralisierten (coiled-coil, CC) Domäne), und RNL (mit einer RPW8 Domäne) zu unterscheiden. Die Diversität der Domänen-Architekturen wird von zusätzlich integrierten Domänen (IDs) noch erhöht. Diese können an unterschiedlichen Stellen im Gen lokalisiert sein. Unterschiedliche Pflanzenarten haben zwischen wenigen Dutzend und mehreren Hundert NLRs. Die intraspezifische Diversität der R Gene ist ebenfalls hoch, und die wenigen bekannten NLRs die für langfristige Resistenzen verantwortlich sind, sind oft spezifisch für einzelne Populationen der Art. Aktuelle intraspezifische NLR Studien weisen mehrere Mängel auf: Das Pan-NLR-om (die Gesamtheit aller NLR Gene und Allele die in einer Spezies vorkommen) kann oft nicht umfassend beschrieben werden. Es werden zu wenige Populationen analysiert und der Nachweis der NLRs wird eigentlich immer durch Referenzgenome gelenkt, was das Finden neuer Gene und Allele erschwert. Zusätzlich dazu können ungeeignete oder unausgereifte bioinformatische Analysepipelines NLRs während des Assemblings oder der Annotationen übersehen, oder in fehlerhaften NLR Annotationen münden. Das Pan-NLR-om einer Pflanzenart zu kennen ist der Schlüssel um neue resistente Pflanzen für die Zukunft zu erhalten. Ich habe eine umfangreiche und verlässliche Datenbank erstellt, die das annähernd vollständige Pan-NLR-om der Modellpflanzenart *Arabidopsis thaliana* definiert. Der Fokus lag auf einem Set von 65 variablen Populationen und dem hochmodernen zielgerichteten Sequenzieren langer DNA Sequenzen (SMRT RenSeq). Meine Analysepipeline wurde so entworfen, dass die optimierten Methoden auch bei jedem anderen SMRT RenSeq Projekt angewendet werden können. Im ersten Teil meiner Thesis lege ich Standards für die Qualitätskontrolle des Assemblings der NLR-kodierenden genomischen Fragmente fest. Des weiteren stelle ich eine neue und sorgfältige Annotationspipeline vor, die von gewissenhafter manueller Reannotation unterstützt wird. Der zweite Teil der Thesis beschreibt das saturierte *A. thaliana* Pan-NLR-om. Die intraspezifische Diversität der NLR Domänen-Architekturen wird gezeigt und der Nutzen von neuen integrierten Domänen hervorgehoben. Der Kern des NLR-oms wird definiert und es wird beschrieben, welche NLRs außerhalb des Kerns Präsenz-Absenz Polymorphismen zeigen. Haplotyp-Sättigung wird gezeigt, Selektion-

skräfte werden quantifiziert und evolutionär gekoppelte koevolvierende NLRs werden detektiert. Die Ergebnisse der Methodenoptimierung zeigen, dass die Qualität eines NLR Assemblings hauptsächlich von der Menge und der Qualität der Sequenzierdaten abhängt. Des weiteren zeigen sie, dass manuelle Reannotation der automatisierten NLR Vorhersage entscheidend ist, um die häufig vorkommenden Misannotationen zu verhindern. Die Sättigung eines NLR-oms wurde bisher in keiner Pflanzenart gezeigt. Diese Studie bietet daher einen neuartigen Blick auf die intraspezifische Varianz der NLRs, des Kerns und der NLR Evolution. NLRs mit integrierte Domänen werden häufiger gefunden als bisher bekannt, was eine zentrale Rolle dieser NLRs in der Interaktion von Pflanze und Pathogen nahelegt. Diese Arbeit legt neue Standards für die Analyse von Genfamilien einer Spezies fest. Zukünftige NLR-om Projekte von wichtigen Nutzpflanzen profitieren von meinen Ergebnissen und der leicht adaptierbaren Analysepipeline. Letzendlich wird dies unser Wissen über die intraspezifische NLR Diversität über die Grenzen weniger Referenzspezies oder Genome hinweg erweitern. Das Auffinden von funktionierenden aktiven NLRs zur Zucht resistenter Pflanzen wird so vereinfacht.

Contents

List of Figures

List of Tables

Glossary

aa-tRNA aminoacyl-tRNA

ACD6 Accelerated Cell Death 6

ADP adenosinediphosphate

ADR1 Activated disease resistance 1

AED Annotation Edit Distance

ATP adenosinetriphosphate

ATR1 *A. thaliana* Recognized 1

Avr-CO39 Avirulence protein Avr-CO39

Avr-Pia Antivirulence protein Avr-Pia

Avr-Pik Antivirulence protein Avr-Pik

AvrAC Type III effector AvrAC

AvrB Avirulence protein B

AvrL567 Avirulence protein AvrL567

AvrPphB Avirulence protein Pseudomonas phaseolicolaB

AvrRpm1 Type III effector AvrRpm1

AvrRps4 Type III effector AvrRps4

BUSCO benchmarking set of universal single copy ortholog

CC coiled-coil

CCS circular consensus sequencing

CLR circular long read

CNL CC-containing NLR

CNV copy number variation

DAMP damage associated molecular pattern

DM2 Dangerous mix 2

DNA desoxyribonucleic acid

dRenSeq diagnostic RenSeq

EF-Tu elongation factor thermo unstable

ENA European Nucleotide Archive

EST expressed sequence tag

ETI effector-triggered immunity

HMA heavy-metal-associated domain

HMM hidden markov model

HopBA1 Type III effector HopBA1

HopF2a Type III effector HopF2a

HopM1 Effector protein HopM1

HopZ1a Type III effector HopZ1a

HP hair pin

HR hypersensitive response

IBS identity by state

ID integrated domain

L5 flax L5 resistance protein

L6 flax L6 resistance protein

L7 flax L7 resistance protein

LRR leucine-rich repeat

M flax rust resistance protein M

MHD methionine-histidine-aspartate

MLA10 MLA10

N tobacco mosaic virus resistance protein N

NB nucleotide-binding

NLP necrosis and ethylene-inducing peptide 1-like protein

NLR nucleotide-binding and leucine-rich repeat containing gene

NRG1 N requirement protein 1

OG orthogroup

ORF open reading frame

PacBio Pacific Biosciences

PAMP microbe- or pathogen associated molecular pattern

PAV presence absence variation

PBL2 Probable serine/threonine-protein kinase PBL2

PBS1 AVRPPHB Susceptible1

PCR polymerase chain reaction

PE paired end

Pik Rice blast resistance gene Pik

Pik-2 Rice blast resistance protein Pik-2

PopP2 Type III effector PopP2

PR pathogenesis-related protein

PRR pattern recognition receptor

PTI PAMP-triggered immunity

QC quality control

QI quality index

QTL quantitative trait loci

R resistance

R8 R8 late blight resistance gene

RBA1 Response to the bacterial type III effector protein HopBA1

RenSeq resistance gene enrichment sequencing

RGA4 Disease resistance protein RGA4

RGA5 Disease resistance gene RGA5

RIN4 RPM1-interacting protein 4

RKS1 G-type lectin S-receptor-like serine/threonine-protein kinase RKS1

RNA ribonucleid acid

RNL RPW8-containing NLR

ROI read of insert

ROS reactive oxygen species

RPM1 Resistance to *Pseudomonas syringae* pv. maculicola 1

RPP1 Recognition of *Peronospora parasitica* 1

RPP13 Resistance to *Peronospora parasitica* 13

RPS4 Resistance to *Pseudomonas syringae* 4

RPS5 Resistance to *Pseudomonas syringae* 5

RPW8 powdery mildew resistance protein, RPW8 domain

RPW8.1 Resistance to Powdery Mildew 8.1

RPW8.2 Resistance to Powdery Mildew 8.2

RRS1 Recognition of *Ralstonia solanacearum* 1

SA salicylic acid

SBS sequencing by synthesis

SMRT single molecule real time

SNC1 Suppressor of NPR1-1, Constitutive 1

SNP single nucleotide polymorphism

SOP standard operating procedure

SQS squalene synthase

SV structural variation

TE transposable element

TIR toll/interleukin-1 receptor homology

TNL TIR-domain containing NLR

Ve1 Verticillium wilt disease resistance gene

WGS whole genome sequencing

WRKY WRKY domain

ZAR1 HopZ-activated resistance 1

ZMW zero mode waveguide

1. Introduction

A substantial amount of the potential yield of crop plants is not realized due to disease. Around 16 % of the annual production is estimated to be lost due to infection with fungi, oomycetes, and bacteria (Oerke 2006) (fig. 1.1 and fig. 1.2). This loss is globally threatening the human food supply. Farmers often must fight yield losses with the application of costly amounts of pesticides and by intensifying their production systems. This drives the fast occurrence and spread of resistant pathogens, and threatens the agricultural sustainability. A better and environmentally friendly way to control plant pests and the associated yield loss is the breeding of resistant plant varieties. Growing these plants increases the crop yield, reduces the need to apply chemical pesticides and allows for a more sustainable use of the agricultural area. Resistant plants are searched in wild populations or related species, and traditional breeding-based methods can be used to introduce resistances into the high-yield domesticated crops.

A thorough understanding of the molecular basis of resistance is needed to create long-term resistant crops. The genes that underlie the resistance and their mode of action need to be identified. The strength of the effect, and the longevity need to be evaluated. Influences on other important traits such as growth and yield need to be taken into account, too. Especially in the light of a fast growing human population that is ultimately depending on increasing amounts of plant food, it is crucial to understand the molecular basis of this plant-pathogen fight. It is of utmost importance to be able to detect successful defense strategies and use those to direct future breeding programs of crops that are needed to secure our food supply.

1.1. The 'zigzag' model of plant-pathogen interactions

Plants are equipped with a diverse set of defense reactions that are induced upon recognition of a pathogenic thread. Those defense reactions are the result of an evolutionary 'arms-race' that lead to the development of a complex innate immunity. According to the 'zigzag' model proposed by Jones et al. (2006), pattern recognition receptors (PRRs) that detect microbe- or pathogen associated molecular patterns (PAMPs) induce the first line of a plant's defense, termed PAMP-triggered immunity (PTI) (reviewed for example in Bigeard et al. (2015), Gust et al. (2017), and Thomma et al. (2011)).

A highly conserved PAMP detected in plants is flagellin (Felix et al. 1999; Gómez-Gómez et al. 2002), which is the major component of the flagellum of motile bacteria. Another detected protein is the elongation factor thermo unstable (EF-Tu), one of the most abundant and conserved proteins in prokaryotes. It is part of the translation

Figure 1.1.: Wheat stem rust disease caused by *Puccinia graminis* by Liang Qu/IAEA, licensed under CC BY 2.0 https://c1.staticflickr.com/3/2839/9685492848_33a85eb8a3_b.jpg

Figure 1.2.: Tomato late blight caused by *Phytophthora infestans* by Scot Nelson https://upload.wikimedia.org/wikipedia/commons/e/e3/Tomato_late_blight_fruit_cluster_%285816739612%29.jpg

machinery of new proteins in the ribosome, where its major function is the transport and the binding of aminoacyl-tRNA (aa-tRNA) to the ribosome. It was shown to be a PAMP in *Arabidopsis thaliana* and other *Brassicaceae* (Kunze 2004). More information about conserved microbe-derived peptides can be found in the review from Albert (2013).

Cell wall components have also been reported to act as PAMPs. In bacteria, peptidoglycan forms a mesh-like structure in the cell walls of both gram-negative and gram-positive bacteria. It was shown to elicit immune responses in *A. thaliana* (Erbs et al. 2008; Liu et al. 2014). Lipopolysaccharide is found in the outer membrane of gram-negative bacteria restricting membrane permeability. Dow et al. (2000) review its role in plant pathogenesis and resistance. Chitin is a fungal cell wall component that was shown to induce immune responses upon perception in tomato (Felix et al. 1993). Glucan is part of the cell walls of fungi and oomycetes. It is recognized for example in tobacco (Klarzynski et al. 2000), rice, and soybean (Yamaguchi et al. 2000).

Because PAMPs are typically widespread in pathogen species, PTI protects against many different pathogens, conferring broad-spectrum resistance. PTI results in the activation of various defense responses intended to restrict pathogen access and multiplication. An example of a strategy that restricts pathogen access to the plant's intracellular space is stomata closure. Stomata are pores mainly found in plant leafs, needed for gas exchange between the plant and its surrounding. In an open state, they provide an easy entry site for pathogens and PTI triggers stomata closure (reviewed in Melotto et al. (2008) and Sawinski et al. (2013)).

Pathogen multiplication is for example hindered by lowering the nutrient supply in the apoplast. Microbes use nutrients involuntarily provided by the plant host (Chen et al. 2010a). Upon silencing of *squalene synthase (SQS)*, Wang et al. (2012) showed increased nutrient efflux into the apoplast when infecting *Nicotiana benthamiana* with *Pseudomonas syringae* and *Xanthomonas campestris*. In the wild type plant, PTI suppressed microbial proliferation by decreasing the transport of nutrients into the apoplast.

Other strategies that limit pathogen proliferation employ by-products of the plant metabolism. Plants normally produce low levels of deleterious reactive oxygen species (ROS) as by-products of the normal oxygen metabolism. Among others, drought, salt-stress, and nutrient deficiency can induce the production of ROS, leading to oxidative stress and significant damage of the plant's cells. Plants also induce the production of ROS as a result to pathogen invasion. This reactive burst is part of the PTI and the toxic features of the reactive oxygen species are helping to remove the pathogenic thread (reviewed in O'Brien et al. (2012)).

PTI also includes the production of phytoalexins. Phytoalexins are antimicrobial secondary metabolites damaging the cell membranes of both bacteria (Rogers et al. 1996) and fungi (Joubert et al. 2011). In *A. thaliana*, the most prevalent phytoalexin is camalexin, which plays a role in resistance to several necrotrophic fungi, to hemibiotrophic oomycetes and fungi, and to powdery mildews (reviewed in Ahuja et al. (2012)). A variety of other defense-related proteins/peptides have been described and are classified into several families of pathogenesis-related proteins (PRs) (reviewed in Loon et al. (2006)).

Pathogens in turn have evolved mechanisms to suppress PTI. They secrete effector proteins into the plant's cells which interfere with PTI-related defense responses. The

bacterial effector Effector protein HopM1 (HopM1) for example actively suppresses the oxidative burst and the stomata closure in *Arabidopsis* and *N. benthamiana* (Lozano-Durán et al. 2014). Another effector, the *P. syringae* protein HopAI1, suppresses the oxidative burst, callose deposition and gene expression that is induced upon recognition of flagellin in *Arabidopsis* (Zhang et al. 2007).

A second line of defense is therefore formed by resistance (R) proteins that are activated by those pathogenic effectors. R proteins induce the effector-triggered immunity (ETI), which often results in a hypersensitive response (HR) leading to programmed local cell death, restricting pathogenic infections to only a small part of the plant (Jones et al. 2006). R proteins are discussed in detail in section 1.2, focusing on their structural composition (section 1.2.1), function (section 1.2.2) and the current knowledge in different species and populations (section 1.3). As the result of an evolutionary arms race between pathogen and plant, effectors are selected for that evade or suppress recognition and ETI. Plant R proteins in turn evolve to overcome those effects and rescue ETI.

The zigzag model has been subject to refinements and extensions and the strict division between PTI and ETI has been shown to be too simplistic (reviewed in Leibman-Markus et al. (2018), Pritchard et al. (2014), and Thomma et al. (2011)). In addition to PAMPs, plants can detect signals from damage associated molecular patterns (DAMPs), molecules from the plant itself that have been changed as a result of the damage caused by pathogens (reviewed in Boller et al. (2009) and Gust et al. (2017)). Cell wall fragments like oligogalaturonides are DAMPs released during microbial infection. They are known to induce PTI responses like the accumulation of phytoalexins, ROS, and others (reviewed in Ferrari (2013)). Cutin is a component of the plant's cuticle that can be split into monomers by fungal enzymes. These monomers are DAMPs that have been shown to induce ROS in cucumber (Fauth et al. 1998). Other known DAMPs include extracellular adenosinetriphosphate (ATP), green leaf volatiles, and several proteins and peptides (Boller et al. 2009; Gust et al. 2017).

In the classical view promoted by the zigzag model, PAMPs are broadly conserved patterns that are important for the pathogen's fitness. They are detected by PRRs, ancient cell surface receptors that are shared between several plant species. Effectors are narrowly occurring only in some species or strains, and are important for pathogen virulence. Effectors are detected from intracellular R proteins that are relatively young. New effectors and R proteins continuously evolve due to an evolutionary arms race between plant and pathogen.

Thomma et al. (2011) argue against the strict division between ETI and PTI and convincingly lay out a variety of examples that show deviations from the classical view. They report conserved effectors and narrowly distributed PAMPs. The cytotoxic necrosis and ethylene-inducing peptide 1-like protein (NLP) effectors for example induce necrotic cell death and various other immune responses (Albert et al. 2015; Böhm et al. 2014; Qutob et al. 2006). They are conserved and widespread among fungi, oomycetes, and bacteria. In contrast, the PAMP Pep-13 is only conserved in *Phytophtora* (Brunner et al. 2002). Some PAMPs or epitopes of the same PAMP, are only detected in certain plant species. The flagellin-derived epitope flg15 for example, is highly active in tomato, but not in *Arabidopsis* or *N. benthamiana* (Meindl et al. 2000; Robatzek et al. 2007).

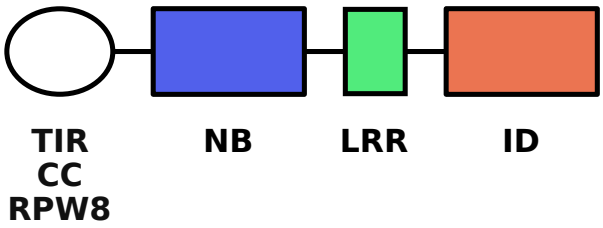


Figure 1.3.: A generic NLR
An NLR typically contains a TIR, CC, or RPW8 domain at its N-terminus, followed by the central NB domain and one or several LRRs. Integrated domains (IDs) might be found at the C-terminus, can also be found at the N-terminus, and more rarely occur between the three canonical domains.

This also suggests that the corresponding PRR is relatively young. Conversely, the R gene *Verticillium wilt disease resistance gene (Ve1)*, mediates resistance in the two fungal species *Verticillium dahliae* and *Verticillium albo-atrum* suggesting a conserved PAMP-like elicitor and PRR-like function of Ve1 (reviewed in Thomma et al. (2011)).

In addition to the increasingly blurry distinction between PAMPs and effectors, PRRs and R proteins, other elements of the zigzag model have been shown to be incomplete. Some PAMPs are reported to be important for the pathogen's virulence in addition to their contribution to microbial fitness. The model also does not include other resistance-shaping events (Pritchard et al. 2014) like environmental factors that may have changed the plant's alertness (e.g. drought, prior exposure to pathogens). Nonetheless, the model provides a good conceptual view of how plants and pathogens interact. Researchers can use the zigzag model as a backbone when asking questions and stay open to extensions and refinements when answering them.

1.2. Nucleotide-binding and leucine-rich repeat containing genes

Among all currently cloned genes acting in PTI or ETI, 61 % are nucleotide-binding and leucine-rich repeat containing genes (NLRs) (Kourelis et al. 2018), which are predominantly responsible for effector detection and signal transduction in ETI. A plant species might contain several hundred NLRs (Shao et al. 2016). NLRs typically contain a central nucleotide-binding (NB) domain and multiple leucine-rich repeats (LRRs) at their C-terminal end. NLRs are classified into three anciently diverged classes by the occurrence of additional N-terminal domains (Shao et al. 2016, 2014; Zhang et al. 2016). TIR-domain containing NLRs (TNLs) possess an additional toll/interleukin-1 receptor homology (TIR) domain, CC-containing NLRs (CNLs) contain coiled-coils (CCs), and RPW8-containing NLRs (RNLs) contain a powdery mildew resistance protein, RPW8 domain (RPW8). Truncated NLRs, whose domains deviate from the typical structural arrangement are widespread (reviewed in Jacob et al. 2013). They may lack the N-terminal domains, the C-terminal LRRs, or even the central NB domain.

Truncated NLRs might still be functional and involved in disease resistance. The *A. thaliana* proteins Resistance to Powdery Mildew 8.1 (RPW8.1) and Resistance to Powdery Mildew 8.2 (RPW8.2) only contain the RPW8 domain and signatures of CCs. They control resistance to a broad spectrum of powdery mildew pathogens (Xiao et al. 2001). The NLR Response to the bacterial type III effector protein HopBA1 (RBA1) (also from *A. thaliana*) contains only a TIR domain and is sufficient to trigger cell death in response to the *P. syringae* effector protein Type III effector HopBA1 (HopBA1) (Nishimura et al. 2017). In addition to the canonical domain set (NB, TIR, CC, RPW8, LRR), NLRs may contain other integrated domains (IDs) with important functions in disease resistance (discussed in section 1.2.2).

1.2.1. Domains

Without a pathogenic trigger, NLRs are normally kept in an 'off-state' mediated by the conserved central NB domain. Mutations in the NB domain, e.g. in the conserved p-loop or methionine-histidine-aspartate (MHD) motif can result in loss of function or autoactivity (Bendahmane et al. 2002; Williams et al. 2011). The binding of adenosinediphosphate (ADP) is thought to secure a closed conformation that prevents signaling, whereas upon pathogen recognition, ATP replaces ADP, which puts the NLR into an active 'on-state' that allows signaling. Evidence for this model was found in a study of the flax rust resistance protein M (M) (Williams et al. 2011). It was shown that autoactivity of M was caused by a mutated MHD motif in the NB domain, which led to preferential binding of ATP.

It is currently proposed that the N-terminal TIR and CC domains act mainly as signaling and oligomerization elements. The CC domain of the Barley MLA10 resistance protein for example forms homodimers in solution and is sufficient for triggering cell death (Maekawa et al. 2011). The TIR domain of the flax L6 resistance protein (L6) self-associates in vitro and forms an autoactive complex (Bernoux et al. 2011). Given that the full L6 protein does not form a complex in vivo, activation by a pathogen effector might be needed for self-association of this NLR. Heterodimerizing NLRs are also known. In *A. thaliana* for example, the TIR domains of the two NLRs *Recognition of Ralstonia solanacearum 1 (RRS1)* and *Resistance to Pseudomonas syringae 4 (RPS4)* form a heterodimeric complex that is required to suppress effector-independent defense signaling of RPS4 (Williams et al. 2014).

The LRR domain is known to inhibit NLR autoactivity and to be involved in effector recognition. The first four LRR repeats of the *A. thaliana* NLR Resistance to *Pseudomonas syringae* 5 (RPS5) inhibit autoactivation in the absence of a pathogenic thread, and the full LRR domain is needed for effector recognition (Qi et al. 2012). Also in *A. thaliana*, the LRR of Recognition of *Peronospora parasitica* 1 (RPP1) mediates the interaction with its corresponding effector *A. thaliana* Recognized 1 (ATR1) from *Hyaloperonospora arabidopsidis* (Krasileva et al. 2010) and the hypersensitive response depends on the TIR domain, in concordance with the reported signaling function of this N-terminal domain.

It is not clear yet, how the RPW8 domain is involved in NLR-mediated resistance. As

already described, the *A. thaliana* NLRs *RPW8.1* and *RPW8.2* are resistance genes only containing the RPW8 domain. They induce localized defense responses when confronted with powdery mildew (Xiao et al. 2005, 2001). RPW8-containing NLRs are known to function as 'helpers', genes necessary for the activation of defense responses after effector-recognition by other 'sensor' NLRs. In *N. benthamiana*, the RPW8-containing N requirement protein 1 (NRG1) is needed for signal transduction upon recognition of the tobacco mosaic virus (TMV) by the NLR tobacco mosaic virus resistance protein N (N) (Peart et al. 2005). The *A. thaliana ADR* clade contains *Activated disease resistance 1 (ADR1)*, *ADR1-L1*, and *ADR1-L2* which code for an N-terminal RPW8 domain and the central NB domain. They are helpers regulating the accumulation of the defense hormone salicylic acid downstream of effector recognition by several other NLRs (Bonardi et al. 2011). Importantly, mutations in the p-loop of the NB domain do not alter the function of ADR1-L2, suggesting that the activation differs from typical TNLs and CNLs, and instead is specific to RNLs. *NRG1*-like and *ADR1*-like genes are present in many higher plants, suggesting a common and conserved biological function (Collier et al. 2011). Recently, RNLs have been hypothesized as signaling hubs in immune receptor networks, which enhance the evolvability of sensor NLRs, and at the same time provide robustness by using conserved sets of signaling elements (Wu et al. 2018).

1.2.2. NLR activation

Diverse NLR activation modes are known. NLRs can be activated by effectors via direct or indirect interaction. In *A. thaliana*, the resistance protein RPP1 associates directly with the ATR1 effector from *H. arabidopsidis* (Krasileva et al. 2010). The effector is recognized by the LRR region of RPP1, and the NB domain is needed for activation (Goritschnig et al. 2016; Steinbrenner et al. 2015). In *Linum usitatissimum*, the NLRs flax L5 resistance protein (L5), L6, and flax L7 resistance protein (L7) directly recognize the Avirulence protein AvrL567 (AvrL567) from the flax rust fungus (*Melampsora lini*). The recognition specificity for different *AvrL567* variants lies within the LRR regions (Ravensdale et al. 2012). Other direct interactions have been found for example in rice (Jia et al. 2000; Maqbool et al. 2015), tobacco (Ueda et al. 2006), potato (Chen et al. 2012), and apple (Meng et al. 2018).

NLRs can detect effector presence indirectly by monitoring other host proteins which are important for the plant's immune response. The NLR guard senses either the binding of an effector to the guardee, or effector-induced modifications. The *A. thaliana* NLR RPS5 detects the cleavage of its guardee AVRPPHB Susceptible1 (PBS1) by the *P. syringae* effector Avirulence protein Pseudomonas phaseolicolaB (AvrPphB) (Qi et al. 2014).

Some proteins - deemed decoys - mimic effector targets. They are guarded by NLRs which induce ETI upon recognition of effector-induced changes in the decoy. The *X. campestris* Type III effector AvrAC (AvrAC) induces molecular changes in the decoy protein Probable serine/threonine-protein kinase PBL2 (PBL2) of *A. thaliana*. The change is perceived by the NLR HopZ-activated resistance 1 (ZAR1) via the intermediate adapter protein G-type lectin S-receptor-like serine/threonine-protein kinase RKS1

(RKS1) (Wang et al. 2015). Also in *A. thaliana*, the NLR Resistance to *Pseudomonas syringae* pv. maculicola 1 (RPM1) guards the RPM1-interacting decoy RPM1-interacting protein 4 (RIN4) (Li et al. 2014a; Mackey et al. 2002, reviewed in Kourelis et al. 2018). Conformational changes of RIN4 are caused by the Avirulence protein B (AvrB) and Type III effector AvrRpm1 (AvrRpm1) of *P. syringae* and are sensed by RPM1, which induces immune responses (Li et al. 2014a). However, both for *RIN4*, and for *PBL2*, research suggested a role in PTI (Kim et al. 2005; Lee et al. 2015; Zhang et al. 2010a), so they might not be real decoys. Other examples for decoys have been found in *A. thaliana* (Zhang et al. 2012) and in tomato (Ntoukakis et al. 2013).

The guard model proposes that NLRs guard important proteins that may be targeted by several pathogens and thus explains how a relatively small set of NLRs can fight the tremendous amount of diverse pathogens that attack plants (Dangl et al. 2001). The *A. thaliana* NLR ZAR1 not only recognizes AvrAC via the decoy PBL2 (Wang et al. 2015), but also recognizes the *P. syringae* Type III effector HopF2a (HopF2a) (Seto et al. 2017) and Type III effector HopZ1a (HopZ1a) (Lewis et al. 2010).

NLRs may contain integrated domains - putative decoys for effector targets - in addition to the canonical domain set (Bailey et al. 2018; Cesari et al. 2014; Kroj et al. 2016; Sarris et al. 2016; Wu et al. 2015). They are thought to have occurred via the duplication of effector target proteins followed by integration into an NLR gene. How they are involved in immunity, and if they act as decoys or retained their original function remains unclear (Kourelis et al. 2018; Sarris et al. 2016; Wu et al. 2015) and needs to be tested for each gene individually. Integrated domains with known effector recognition capability are known from the *A. thaliana* TNL *RRS1* (Le Roux et al. 2015; Sarris et al. 2015) and the rice CNLs *Disease resistance gene RGA5 (RGA5)* (Cesari et al. 2013; Ortiz et al. 2017) and *Rice blast resistance gene Pik (Pik)* (Maqbool et al. 2015). RRS1 contains a WRKY domain (WRKY) that can directly bind to the *P. syringae* Type III effector AvrRps4 (AvrRps4) (Sarris et al. 2015). The WRKY domain can also be acetylated by the *Ralstonia solanacearum* Type III effector PopP2 (PopP2) (Le Roux et al. 2015). The direct interaction with the effector AvrRps4 triggers immune responses depending on the NLR RPS4. The acetylation by PopP2 prevents recognition of AvrRps4 in the accession Col-0, but not in Nd-1 and Ws-2 (Le Roux et al. 2015; Sarris et al. 2015). *RGA5* and *Pik* both contain an integrated heavy-metal-associated domain (HMA) (Cesari et al. 2013; Maqbool et al. 2015; Ortiz et al. 2017). RGA5 directly binds to the *Magnaporthe oryzae* effectors Antivirulence protein Avr-Pia (Avr-Pia) and Avirulence protein Avr-CO39 (Avr-CO39) and the immune response is mediated via the NLR Disease resistance protein RGA4 (RGA4). Pik directly recognizes Antivirulence protein Avr-Pik (Avr-Pik) (also from *M. oryzae*) and the immune response is depending additionally on the NLR Rice blast resistance protein Pik-2 (Pik-2).

All three known NLRs with effector-recognizing integrated domains are genomically paired with the respective canonical NLR needed for signal transduction. Genomically paired NLRs are frequent in plants (Kroj et al. 2016; Narusaka et al. 2009; Stein et al. 2018) and are hypothesized to be also functional pairs, especially when found in head-to-head orientation.

1.3. Species- and population-wide analyses of NLRs

While functional analyses elucidate individual NLR-pathogen interactions, large-scale comparisons between or within species allow characterizing NLR complements, diversity and evolution.

1.3.1. Interspecific NLR studies

The total number of reported NLRs fluctuates highly between different plants (Shao et al. 2016; Zhang et al. 2016). Around 155 NLRs are reported in *A. thaliana* (149 in Meyers et al. (2003), 159 in Guo et al. (2011)), while important crop plants like tomato (355 NLRs, Andolfo et al. (2014)), potato (438 NLRs, Jupe et al. (2012)), and rice (508 NLRs in Li et al. (2010a), 535 NLRs in Stein et al. (2018)) contain up to three times as many NLRs. NLR numbers must not be confused with defense capability. *Cucumis sativus* for example contains 57 NLRs (Wan et al. 2013) and the NLR complement of *Carica papaya* contains only 54 genes (Porter et al. 2009).

Studies with overlapping species sets almost always report varying NLR numbers. Genomes are often first published in a draft state and are updated when new data allows for significant improvements in completeness and continuity. NLRs often sit in genomic regions that are hard-to assemble and result in non-continuous draft assemblies. A new and improved version of a species' genome thus often changes the view of the species' NLR complement. In addition, annotation methods can differ greatly since there is no standard gene annotation pipeline in the field and there is also no universally accepted consensus about which domains need to be present to define an NLR. All reported NLR gene numbers thus have to be seen in the light of the applied methods and input datasets. A survey of NLR gene numbers in 22 angiosperm species (for phylogenetic relations see fig. 1.4) comparing the results from 12 different papers is given in table 1.1. The species set is drawn from Shao et al. (2016), who analyzed evolutionary patterns and defined three anciently diverged NLR classes (TNLs, CNLs, and RNLs) in those angiosperms. The NLR numbers are compared to 11 other papers (publication dates ranging from 2003 to 2016). The reported range might guide an estimation of the actual NLR complement of each species. A bigger NLR survey including more species and papers can be found in table A.3.

Not only the total number of NLR genes shows a high variability, but also their distribution in the three anciently diverged classes TNLs, CNLs, and RNLs (Gao et al. 2018; Shao et al. 2016). TNLs are expanded in the Brassicaceae, but vanished completely from the Poaceae (Shao et al. 2016; Yue et al. 2012). CNLs are the ancestral NLR class and dominate in many plants, and RNLs are a basal monophyletic group that is present with rather low numbers in plant genomes (Gao et al. 2018; Shao et al. 2016). Expansions and contractions influenced the establishment of the NLR complements (Shao et al. 2016; Yue et al. 2012; Zhang et al. 2014; Zheng et al. 2016) via tandem duplication, ectopic and segmental duplication, as well as polyploidization and gene losses (Guo et al. 2011; Hofberger et al. 2014; Leister 2004; Plomion et al. 2018; Zhang et al. 2014; Zheng et al. 2016). Transposable elements (TEs) are often associated to NLR loci and may be

Table 1.1.: NLR complements of 22 Angiosperms
Survey of known NLR gene complements for 22 angiosperm species analyzed in Shao et al. (2016). NLR numbers from 12 papers are reported: Shao et al. (2016) (1), Zhang et al. (2016) (2), Zheng et al. (2016) (3), Sarris et al. (2016) (4), Shao et al. (2014) (5), Peele et al. (2014) (6), Yu et al. (2014) (7), Andolfo et al. (2014) (8), Kim et al. (2012) (9), Jupe et al. (2012) (10), Guo et al. (2011) (11), Meyers et al. (2003) (12). For an extended list of analyzed species see table A.3.

Species	NLRs	Ref.	Species	NLRs	Ref.
Amborella trichopoda	105	1	*Medicago truncatula*	571	5
Amborella trichopoda	88	2	*Medicago truncatula*	571	1
Arabidopsis lyrata	241	9	*Medicago truncatula*	771	2
Arabidopsis lyrata	134	6	*Medicago truncatula*	770	3
Arabidopsis lyrata	204	4	*Musa acuminata*	111	1
Arabidopsis lyrata	198	1	*Musa acuminata*	105	2
Arabidopsis lyrata	202	2	*Oryza sativa*	595	4
Arabidopsis lyrata	185	11	*Oryza sativa*	498	1
Arabidopsis thaliana	238	9	*Oryza sativa*	470	2
Arabidopsis thaliana	135	6	*Oryza sativa indica*	616	9
Arabidopsis thaliana	165	1	*Oryza sativa japonica*	578	9
Arabidopsis thaliana	213	7	*Phaseolus vulgaris*	406	4
Arabidopsis thaliana	168	2	*Phaseolus vulgaris*	337	5
Arabidopsis thaliana	213	4	*Phaseolus vulgaris*	337	1
Arabidopsis thaliana	149	12	*Phaseolus vulgaris*	334	2
Arabidopsis thaliana	149	11	*Phaseolus vulgaris*	359	3
Brachypodium distachyon	185	9	*Phyllostachys heterocycla*	344	1
Brachypodium distachyon	501	4	*Sesamum indicum*	170	1
Brachypodium distachyon	253	1	*Setaria italica*	470	4
Brachypodium distachyon	327	2	*Setaria italica*	424	1
Brassica rapa	204	1	*Setaria italica*	380	2
Brassica rapa	248	7	*Solanum lycopersicum*	264	4
Brassica rapa	151	6	*Solanum lycopersicum*	255	1
Brassica rapa	207	4	*Solanum lycopersicum*	223	2
Brassica rapa	196	2	*Solanum lycopersicum*	355	8
Cajanus cajan	256	2	*Solanum tuberosum*	543	4
Cajanus cajan	289	5	*Solanum tuberosum*	447	1
Cajanus cajan	289	1	*Solanum tuberosum*	355	2
Cajanus cajan	815	3	*Solanum tuberosum*	438	10
Capsella rubella	75	6	*Sorghum bicolor*	317	9
Capsella rubella	152	4	*Sorghum bicolor*	422	4
Capsella rubella	127	1	*Sorghum bicolor*	326	1
Capsella rubella	131	2	*Sorghum bicolor*	310	2
Capsicum annuum	305	1	*Thellungiella salsuginea*	88	1
Capsicum annuum	661	2	*Vitis vinifera*	590	9
Glycine max	325	9	*Vitis vinifera*	323	4
Glycine max	784	4	*Vitis vinifera*	545	7
Glycine max	465	5	*Vitis vinifera*	295	2
Glycine max	465	1	*Vitis vinifera*	314	1
Glycine max	442	2	*Vitis vinifera*	754	3
Glycine max	744	3	*Zea mays*	122	9
Medicago truncatula	668	9	*Zea mays*	191	4
Medicago truncatula	1074	4	*Zea mays*	139	1
Medicago truncatula	753	7	*Zea mays*	129	2

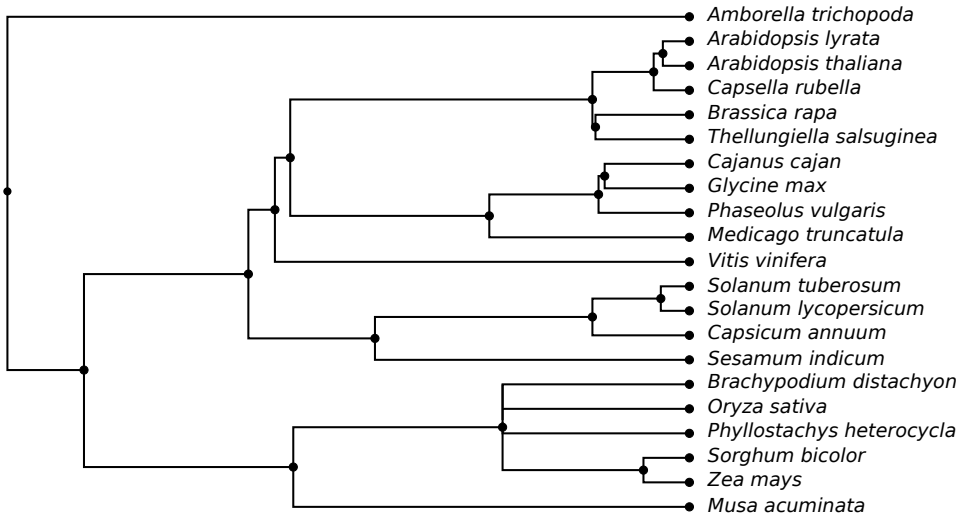


Figure 1.4.: Phylogeny of 22 angiosperms
Visualization of the phylogeny of angiosperm species analyzed in Shao et al. (2016) and surveyed in table 1.1. Phylogeny and tree visualization are based on TimeTree (Kumar et al. 2017, *Setaria italica* is not in the database and thus excluded).

involved in NLR duplication or transposition, too (Henk et al. 1999; Kawakatsu et al. 2016; Kim et al. 2017).

Interspecific comparisons of the NLR complements from 22 angiosperm species resulted in the phylogenetic reconstruction of a set of only 23 ancestral NLR genes from the three classes TNL, CNL, and RNL (Shao et al. 2016). Going even further back, the origin and early diversification has been tackled recently (Gao et al. 2018, reviewed in Ortiz et al. 2018) and it has been shown that NLRs likely first appeared in the charophytes (Gao et al. 2018).

1.3.2. Intraspecific NLR studies

Different populations can respond with contrasting phenotypes to the same pathogen. The reason for those differences are genetic and novel sources of disease resistance can be discovered by defining the responsible NLRs. Presence/Absence polymorphisms of RPM1, for example explain the resistance or susceptibility of *A. thaliana* to the *P. syringae* pv. tomato DC3000 carrying AvrRpm1 (Grant et al. 1998). Allelic diversity at the RRS1 locus in *A. thaliana* justifies why Nd-1 is resistant and a Col-0 derivative is susceptible to *R. solanacearum* (Deslandes et al. 2003). Similarly, allelic diversity between the accessions Ag-0 and Col-0 at the RBA1 locus explains the resistance/susceptibility phenotype to *P. syringae* carrying HopBA1 (Nishimura et al. 2017).

Intraspecific population-based studies have the potential to increase the resolution of NLR variation beyond what is known from reference genomes. This was addressed recently in a study of nine *Brassica oleracea* varieties (Golicz et al. 2016). More than 40 % of the NLR genes that differed between the tested varieties were found in the non-reference aligned and *de novo* assembled portion of the pan genome. A similar

observation was made using different accessions of the model crucifer *A. thaliana*. NLR-containing fragments did not align completely to the reference (Cao et al. 2011). More evidence for the shortcoming of reference-centered NLR analyses in *A. thaliana* comes from the *Dangerous mix 2 (DM2)* cluster, which contains two *RPP1*-like genes in Col-0, and up to seven in Ler (Chae et al. 2014; Stuttmann et al. 2016).

Population-based studies revealed allelic and structural variation (SV) of NLRs. Among the most prominent findings were presence absence variation (PAV) and copy number variation (CNV). A genome-wide comparison of NLRs from 80 *A. thaliana* accessions showed more PAV in NLRs than in the genomic average, and PAV being more often found in clustered NLRs than in singletons (Guo et al. 2011). A study in rice (Yang et al. 2006) also reported variation mainly in clustered NLRs. CNV was found for 33 % of the *A. thaliana* NLRs compared to 12.5 % in the remaining genome (Guo et al. 2011). NLRs were enriched in genes showing SV in *A. thaliana* (Kawakatsu et al. 2016) and in genes showing CNV in *Glycine soja* (Li et al. 2014b), and several genes related to disease resistance were found in PAV genes in *B. oleracea* (Golicz et al. 2016). Allelic variation is reflected in many different haplotypes that are found across NLR loci, e.g. in *A. thaliana* (Cao et al. 2011) or *Malus domestica* (Duan et al. 2017). Some NLRs in *A. thaliana* are recombination hotspots with increased meiotic crossovers (Choi et al. 2016), or have revealed allelic series (Rose et al. 2004).

Consistent with these findings, signs for balancing selection were found e.g. in *A. thaliana* (Noel 1999; Tian et al. 2002). A study of the LRR domains of 27 NLRs in 96 *A. thaliana* accessions demonstrated evidence of balancing selection for some NLRs including *Resistance to Peronospora parasitica 13 (RPP13)*, and provided evidence for selective and partial selective sweeps (Bakker et al. 2006). Especially prone for diversifying selection is the LRR domain of NLR genes (Bakker et al. 2006; Chen et al. 2010b; Kuang et al. 2004; Mondragón-Palomino et al. 2002) in concordance with its suggested role in effector recognition. Intraspecific studies found signs for purifying selection in NLR genes, too. Purifying selection dominated in an evolutionary analysis of five *Rosaceae* species (Zhong et al. 2015) and in the legume family (Zheng et al. 2016), and was found for some NLRs in *A. thaliana* (Bakker et al. 2006) and *Lactuca sativa* (Kuang et al. 2004). Different evolutionary trajectories have been found for NLR genes, diversifying and purifying selection as well as balancing selection reflect the past and ongoing struggle between plants and pathogens (reviewed in Jacob et al. 2013; McDowell et al. 2006).

1.4. Technological limitations

As discussed above, many valuable findings were drawn from population based studies in several plant species. Still, the number of those studies is scarce compared to the many published interspecific comparisons which focus on one reference per species. Population-based studies of the whole pan-NLR'ome of a species are even rarer. Studies often focused only on a specific part of the NLR complement, not necessarily on the complete genes (Bakker et al. 2006), or used only a limited number of accessions (Golicz et al. 2016;

Stam et al. 2016; Zhang et al. 2010b) that did not cover the species' diversity and thus could not provide a comprehensive overview of the pan-NLR'ome, either.

1.4.1. Next generation short read sequencing with Illumina

A short introduction into next generation short read sequencing is needed to understand the limits of population-based NLR studies and possible solutions. In 2015, up to 90 % of the world's sequencing data were produced with Illumina's next generation sequencing by synthesis (SBS) method (Illumina 2015), and the trend is rising. The advantages of SBS are the high amounts of sequencing data that can be produced by parallel sequencing and the price, which is cheap compared to the first generation sequencing methods like Sanger. Sequenced reads can be up to 300 bp long (MiSeq machine).

The method detects single bases as they are incorporated into an extending desoxyribonucleic acid (DNA) strand. A library of DNA fragments is bound to the lanes in a flow cell and each fragment is clonally amplified to form a cluster of identical pieces of DNA. For each single stranded DNA cluster, the complementary strand is synthesized in a controlled base-per-base manner. In repeated synthesis cycles, all four nucleotides tagged with different fluorescent dyes are competing to be added to the extending complementary DNA strand. The correct nucleotide is determined by the sequence of the template cluster and is covalently bound to extend the synthesized DNA. After the binding, the flourescent signal which reflects the incorporated base is detected. The number of cycles with nucleotide binding and flourescence detection determines the length and the sequence of the synthesized DNA read. Upon completion of sequencing the first read, the method allows for sequencing a second read starting from the other end of the clustered DNA template. SBS creates millions of reads representing all the fragments of the input DNA library.

Even though the sequencing accuracy at every position is very high (less than 1 % error, Quail et al. (2012) and Schirmer et al. (2016)), sequencing errors scale with the size of the sequenced genome or genomic proportion. If sequenced only once, these errors cannot be distinguished from true single nucleotide polymorphisms (SNPs) or hinder the correct positioning of reads in the genome. Sequencing the same input DNA several times increases the coverage (or read depth), and can resolve the problem. Each position in the input DNA is then covered by several independently created reads, compensating the small amount of errors that should occur randomly in the genome.

Depending on the underlying research question, the sequenced reads can be mapped back to a known reference genome, or can be assembled *de novo* without the need of further genomic information.

1.4.2. Limitations of short read based NLR studies

The NLR'ome of a plant contains many clustered NLRs with high copy numbers and highly repetitive coding sequences, especially in tandem repeats. The correct assembly of those NLR clusters can be hampered in short-read based assemblies (Witek et al. 2016a). In regions containing highly similar NLRs, short reads sometimes do not provide enough

SNP information to distinguish between neighboring genes. As a result, the assembly contains one gene representing several or all NLRs from the cluster. Increasing the read depth may rescue some of the incorrect merges of repetitive NLRs, but especially in larger genomes, this comes with a significant cost increase. In addition, other genomic features like transposable elements or repetitive elements are known to be found in close proximity to NLRs or in their introns (Henk et al. 1999; Kim et al. 2017). These further hinder the correct assembly of NLR genes when using short reads. With a fixed budget, cost-benefit calculations determine if a project profits more from focusing on a correct assembly, or from including a bigger number of accessions.

1.5. Technological innovations

Today, pan-NLR'ome studies can benefit from technological and methodological advances. Sequencing methods are continuously being developed further especially focused on increasing the read length (section 1.5.2). Elaborate filtering of genomic DNA prior to library preparation, allows for targeted sequencing of genomic subsets, e.g. NLRs (section 1.5.1).

1.5.1. Resistance gene enrichment sequencing (RenSeq)

Instead of sequencing the full genome of each accession, resistance gene enrichment sequencing (RenSeq) (Jupe et al. 2014) focuses on the NLR containing subset of the genome. This simplifies the assembly task by reducing the size and the complexity of the genomic sequence that needs to be assembled, and provides a high read-depth (Gasc et al. 2016; Jupe et al. 2013) for the NLR gene complement. RenSeq uses custom sets of short (120 bp) biotinylated ribonucleid acid (RNA) baits created from known NLR genes in plants related to the study target. The baits have the capacity to bind to DNA fragments with at least 80 % sequence identity (Jupe et al. 2013), which allows to capture the unknown NLR complement in the targeted plant's genome. The workflow is shown in fig. 1.5. A library of single stranded DNA is mixed with the biotinylated baits. The baits bind to DNA fragments containing NLRs. The DNA-bait complex is captured by magnetic beads, which allows separation of the NLR-containing DNA from the rest using simple magnetic force. Beads and baits are then removed. The captured fragments are multiplied by polymerase chain reaction (PCR) prior to sequencing, which provides the desired high coverage. RenSeq has first been successfully applied to the NLR'ome of *Solanum tuberosum* (Jupe et al. 2013), increasing the number of known potato NLRs from 438 to 755. In *Solanum lycopersicum*, RenSeq was used to identify 105 novel NLRs, and 126 NLRs were found new in *Solanum pimpinellifolium* (Andolfo et al. 2014). In addition, 25 % of the previously described tomato NLR complement could be corrected (Andolfo et al. 2014). Stam et al. (2016) used RenSeq to analyze polymorphisms and evolutionary pressures in the wild tomato species *Solanum pennellii* and found 13 NLRs at which polymorphisms were maintained in the population. It was also used to identify molecular markers that co-segregate with resistance to *Phytophtora*

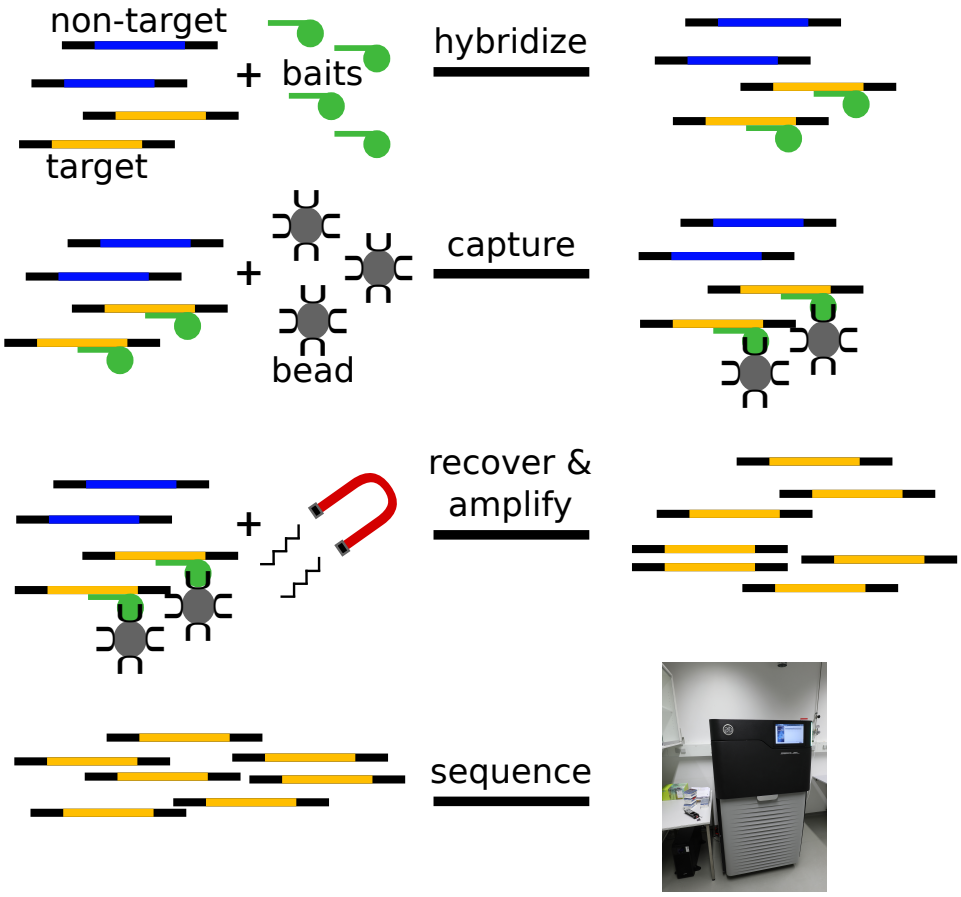


Figure 1.5.: Resistance gene enrichment sequencing (RenSeq) workflow
A single stranded DNA library containing adapters (black), non-target regions (blue), and target regions (orange), is mixed with the bait set (green). The baits hybridize with target regions. Target-bait complexes can be captured and recovered with magnetic beads and are amplified prior to sequencing, here exemplified with PacBio's Sequel machine.

infestans in species without an available reference genome (Jupe et al. 2013). Diagnostic RenSeq (dRenSeq) can be used to identify functional NLRs and validate their sequence. It was applied in Van Weymers et al. (2016) to identify a resistance gene against *P. infestans* in the wild potato *Solanum okadae*. It was used in Jiang et al. (2018) to further specify field resistance against *P. infestans* from the known quantitative trait loci (QTL) *dPI09c* to the NLR *R8 late blight resistance gene (R8)*. Armstrong et al. (2018) provide evidence that dRenSeq can generally be used to identify known functional NLRs to several pathogens in potato varieties. Compared to whole genome sequencing (WGS), RenSeq cost-efficiently provides the high read-depth needed to detect integrity or SNPs, also in the surrounding regions.

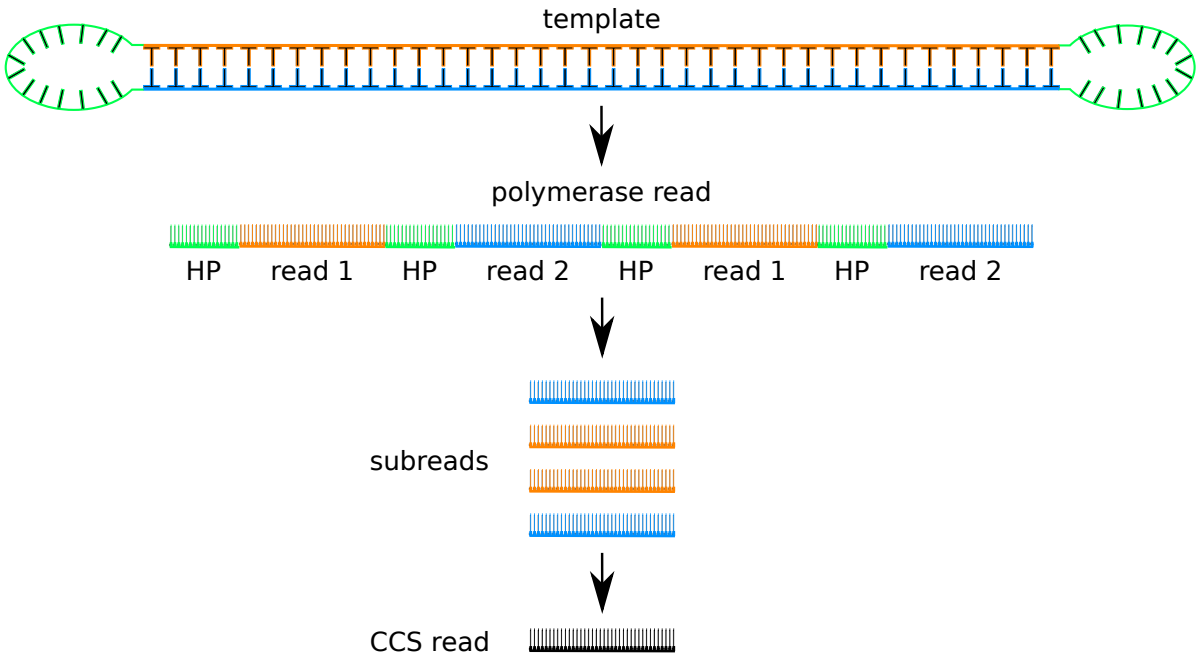


Figure 1.6.: Circular Consensus Sequencing (CCS)
Scheme of circular consensus sequencing (CCS). The double stranded template is shown. The forward read 1 (orange) and the reverse read 2 (blue) are connected via a hairpin adapter (HP, green) at each end. The polymerase produces a long polymerase read that contains several subreads (read 1 and read 2), divided by the HP. The subreads are stacked and a consensus is formed (CCS read).

1.5.2. Single molecule real time (SMRT) sequencing

A relatively recently developed long read sequencing method is the single molecule real time (SMRT) sequencing technology from Pacific Biosciences (PacBio). It is used to assemble high quality *de novo* genomes (Jiao et al. 2017; Teh et al. 2017), spanning repetitive and complex regions (Bao et al. 2017) and resolving structural variants (Huddleston et al. 2016).

SMRT sequencing detects the incorporation of **S**ingle nucleotide **M**olecules in **R**eal **T**ime. The DNA is sequenced on a SMRT cell that contains tens of thousands of zero mode waveguides (ZMWs), extremely small holes in which single DNA fragments can be sequenced. The prepared DNA library contains long (several thousand bases) DNA-polymerase complexes that bind to the bottom of ZMWs (one per ZMW). The surrounding medium contains the four nucleotides labeled with different flourescent dyes. As in Illumina's short read sequencing method, the incorporation of different nucleotides produces colored light signals that in SMRT sequencing are detected per molecule at the bottom of each ZMW and do not require a pause between steps. The light pattern is translated into the nucleotide sequence of the sequenced read. The method is fast, and read lengths of more than 100 000 bp can be obtained.

A useful feature of SMRT sequencing is that the template DNA fragment is double

stranded, with a hair pin (HP) adapter at each end (fig. 1.6). Sequencing does not stop when the template's forward strand is read completely. Instead, the polymerase continues to alternately read the reverse and forward strand until its life time is over. The template's length and the polymerase's life time determines if and how often forward and reverse strands are read. The standard circular long read (CLR) sequencing produces one long sequence from the forward strand of the template molecule. The template fragment is here too long to be read completely, and the sequencing length is solely determined by the life time of the polymerase. The polymerase read shown in fig. 1.6, would here only contain the first HP, and parts of the read 1. CCS sequencing creates multiple passes for each template molecule. Templates are shorter than for CLR, allowing the polymerase to continue reading at the end of read 1.

CCS sequencing produces the completely sequenced forward strand (read 1 in fig. 1.6) followed by the sequence of the hairpin adapter and then - alternating - reverse strand (read 2 in fig. 1.6), hairpin, forward strand, hairpin, until the polymerase stops or falls off the template (polymerase read in fig. 1.6). Compared to Illumina's short read sequencing technology, SMRT sequencing produces more per-base errors, on the order of 10-15% (Korlach 2013) instead of less than 1%, but CCS sequencing provides a direct internal error correction step. Forward and reverse reads from the same DNA template are stacked and one consensus read (CCS read) is produced. The consensus read contains less errors, 15 passes have been reported to secure 99 % read accuracy (Eid et al. 2009).

Sequencing costs dropped drastically in the last years and long read sequencing became affordable even when large amounts of sequence data are needed. Long reads are easier to assemble than short reads and long read based assemblies are more continuous, especially in non-trivial genomic regions. Longer reads have more overlaps with other reads, which simplifies the search for their correct position. Repetitive genomic regions, like transposons or duplicated genes, cannot be spanned by short reads and result in a high assembly fragmentation. Long reads in turn can span those repetitive regions and thus increase the continuity of the assembly.

The combination of RenSeq with long reads from PacBio SMRT sequencing (SMRT RenSeq) has the potential for assembling unfragmented full NLR genes, NLR gene clusters, and the surrounding regulatory elements (Giolai et al. 2016). It already proved useful in identifying resistance genes against *P. infestans* in the tomato and potato relative *Solanum americanum* (Giolai et al. 2017; Witek et al. 2016b). Witek et al. (2016b) also showed its superiority to short read RenSeq data.

1.6. The *A. thaliana* pan-NLR'ome

The work presented in this thesis was carried out as part of a larger international collaboration that aimed to investigate inter- and intraspecific NLRs in important crop plants and their wild relatives using RenSeq. The 'Resistance Gene Diversity' project of 'The 2Blades Foundation' is focused on defining important building blocks of the plant immune system. The investigated species come from three plant families: The Brassicaceae which include the important model species *A. thaliana*, cabbage, mustard

greens, turnips, canola, and relatives, will be used to establish how to best analyze more complex genomes. The Solanaceae contain for example important crops like potatoes and tomatoes and the Triticeae include for example wheat, barley, and related grains.

The objective of this work was the dissection of the full NLR'ome in the model species *A. thaliana*, and the analysis of the population wide intraspecific variability of this important gene family.

The method optimization chapter answers important questions about best-practice techniques to be used when processing SMRT RenSeq data (chapter 2). It shows the minimal requirements for amounts and quality of the input read data. Then it analyzes how to best assemble the read data to produce a continuous and correct representation of the NLR-containing regions of the genome. Subsequently, gene and NLR annotation methods are optimized with a focus on preventing wrongly fused or merged genes, and on correctly predicting domains. Finally, it reports the development of an evaluation of the Quality and the Completeness of accession-specific NLR complements that is needed to ensure that the questions answered in chapter 3 are based on reliable ground assumptions. The NLR'ome is defined comprehensively using 65 diverse *A. thaliana* accessions (chapter 3). SMRT RenSeq is used and automated assembly and annotation methods are combined with manual curation to optimally annotate NLRs as reported in chapter 2. Novelties on several levels are reported. The architectural diversity is described, and novel integrated domains and architectures are analyzed. The NLR'ome shows saturation in the dataset, which allows defining core NLRs, as well as presence-absence polymorphisms in accessory (shell) NLRs, and NLRs only present in few accessions (cloud). Nucleotides and haplotypes show saturation, too. Selective forces that act on NLRs, domains and positions are studied and co-evolving NLRs are described.

2. Method Optimization and Quality Control

In this chapter I report the method optimization steps that were needed to portray the NLR complements of 65 *A. thaliana* accessions comprehensively and correctly. I worked with a set of 73 accessions that represented the species diversity. For these accessions, Resistance gene enrichment and long read sequencing (SMRT RenSeq, Giolai et al. (2016)) data were produced (1/3 by me, 2/3 in collaborating labs, table A.2). I needed to assemble the captured DNA fragments correctly into larger genomic regions (contigs), genes needed to be annotated on those contigs, and NLRs needed to be defined based on their specific domain content. Following assembly and annotation, these genes allowed the dissection of the full NLR repertoire and its variability in the model species *A. thaliana* (chapter 3).

2.1. Assembly Optimization

The basis for a reliable NLR gene annotation is the contiguously and correctly assembled genomic region of origin. Thus, optimizing the individual accessions' assemblies was a crucial task of my PhD project. The best possible assembly is the result of the perfect interplay between the assembly program, the chosen assembly parameters, and the input data. I analyzed the influence of the input data and fine-tuned the different assembly parameters for the long-read assembler `Canu` (version1.3, Koren et al. 2017).

2.1.1. Assembly with `Canu`

`Canu` is a single-molecule assembler written to overcome difficulties like incorrect repeat separation and low per-base coverage. NLRs often occur in clusters of several similar or even identical genes. Correct repeat separation was thus a crucial task for the creation of a complete NLR complement. SMRT RenSeq results in highly covered NLR genes, but only extends with low coverage into the intergenic or non-NLR gene regions in between. The correct assembly of those lowly covered regions was thus also essential. `Canu` was chosen because it tackles these key features of the input RenSeq datasets.

`Canu` first detects overlaps in the input sequences, then generates corrected consensus sequences using those overlaps. Afterwards it trims the consensus sequences and assembles the reads into contigs. For PacBio's corrected CCS reads (section 1.5.2; further only called 'reads'), consensus sequences do not need to be generated and this step

Table 2.1.: Assembly statistics
Assembly statistics for six representative values of the 'genomesize' parameter using the Col-0 RenSeq dataset.

genomesize [Mb]	1	1.5	2	2.5	5	10
Contigs	170	170	170	170	170	170
Length [bp]	1 806 346	1 806 346	1 806 345	1 806 344	1 806 346	1 806 353
Misassembled contigs length [bp]	66 893	66 893	66 893	66 893	66 893	66 901
Mismatches per 100 kbp	12.76	12.76	12.76	12.76	12.76	12.76
Indels per 100 kbp	18.23	18.23	18.29	18.29	18.23	18.23
Genes [full+partial]	157 + 8	157 + 8	157 + 8	157 + 8	157 + 8	157 + 8

can be omitted. `Canu` provides several parameters that can be adjusted to optimize the assembly considering the input features.

2.1.1.1. Optimal choice of the 'genomesize' parameter

The genome size parameter 'genomesize' is an estimate that is influencing the read correction and the overlapping sensitivity. It needs to be set roughly to the expected assembly size. A RenSeq assembly is expected to contain all NLR genes fully assembled, and flanking regions left and right of each NLR depending on the input fragment size. The targeted fragment size was set to 3 kb. Fragments containing the border of an NLR gene and extending into the neighboring non-NLR regions thus provided the necessary information to assemble ∼3 kb flanking regions neighboring each NLR gene. The total expected assembly size from the reference Col-0 was ∼1.74 Mb (168 NLR genes containing 730 kb and 6 kb flanking region per gene). I tested for the Col-0 RenSeq dataset values for the 'genomesize' parameter between 1 Mb and 10 Mb (table 2.1). The resulting assembly size only marginally changed and assembly quality was comparable (`Quast` based comparison to the reference genome). Thus I set 'genomesize=2M' for all further analyses.

2.1.1.2. Influence of the 'errorRate' parameter

The 'errorRate' parameter is the expected error of a single read. It influences which overlaps are generated, and how they are filtered before trimming and unitig construction. During read correction, it defines the maximum errors. In the RenSeq datasets,

reads contained at most 10 % errors (Quality=10, calculated with eq. (2.1), see also CCS calling in section 1.5.2). I tested the 'errorRate' parameter ($0.1 - 0.001$) for the RenSeq Col-0 dataset using different minimum input read quality cutoffs ($\geq$Q10 - $\geq$Q40 equivalent to max 10 % - 0.01 % error per read). I evaluated the percentage of fully covered Col-0 NLR genes for each combination of the 'errorRate' and minimum read quality. Intermediate settings for the 'errorRate' resulted in high rates of NLR gene coverage (fig. 2.1) independent of the input data quality. I chose the setting 'errorRate=0.01' in all subsequent assemblies, to secure a high NLR gene coverage independent from fluctuations in the input read quality of different accessions. As a side note, newer versions of `Canu` (starting v1.5) do not require to set the 'errorRate' any more (Phillippy et al. 2018).

2.1.1.3. Other useful parameter settings

I used the parameter 'pacbio-corrected' to suppress `Canu`'s internal read correction step. Read correction is only needed in an assembly that uses PacBio subreads, whereas an assembly with pre-corrected CCS reads does not benefit from an additional read correction step. `Canu` verifies correctly sequenced positions by creating read overlaps. The parameter 'trimReadsCoverage=X' specifies how many overlaps are needed. Each position in a read that can not be overlapped with at least X other reads is trimmed. Strict overlap-based trimming improved the assembly quality (less mismatches, InDels and misassemblies) in Col-0 tests at the cost of the assembly size and the NLR coverage. Thus, the minimum read depth needed for each position was set to 'trimReadsCoverage=2' which resulted in only two misassembled contigs of which none contained an NLR. `Canu`'s parameters were optimized to correctly assemble large and contiguous contigs that contain full-length NLRs. The impact of the input data was analyzed using those parameters.

2.1.2. Influence of the input data on the assembly

The input data influences an assembly by its quantity (total bases) and quality (read accuracy and read length). A high amount of sequenced bases results in a big and contiguous assembly because of the high per-base coverage. However, if the sequenced reads contain many low-quality bases, misassemblies and wrong base calls accumulate. Alternatively, filtering reads for a high base-calling accuracy secures the assembly accuracy, but assembly contiguity might suffer from the lowered per-base coverage. Longer reads generally improve the contiguity of an assembly. I analyzed how to balance between quality and quantity in order to get the best possible assembly. I specifically tested how the assembly size, quality, and the number of NLRs changed with varying input amount and quality using four RenSeq datasets with complete whole-genome-sequencing based high quality reference genomes (Col-0, Ty-1, KBS-Mac74, and Cvi-0; reference genomes unpublished, yet).

Each corrected CCS read (see section 1.5.2) had a minimum of 90 % per base accuracy (Q10). Sequence numbers and contained bases decreased with increasing minimum

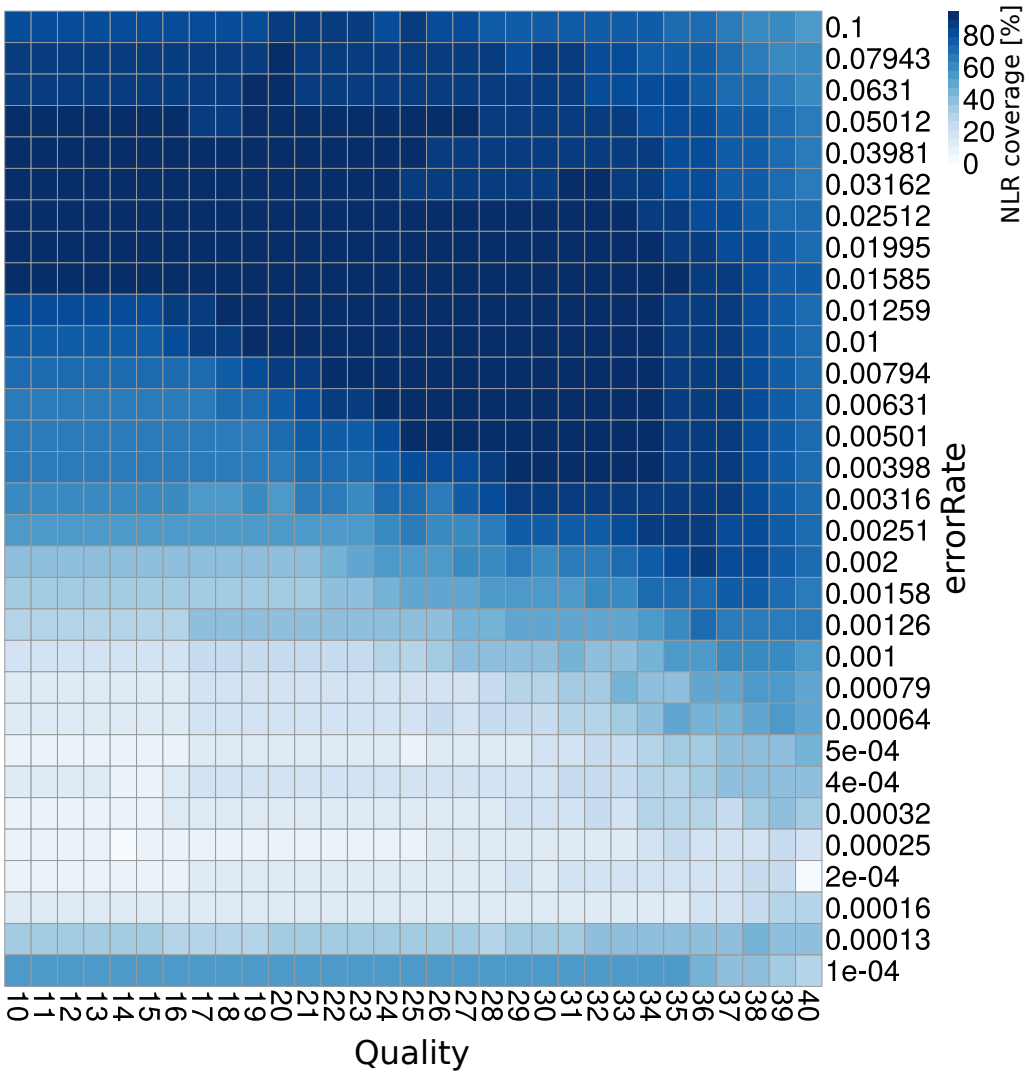


Figure 2.1.: NLR coverage depending on read Quality and errorRate
NLR coverage in dependence of minimum input read 'Quality' and the 'errorRate' parameter of `Canu` using the RenSeq Col-0 dataset. The NLR coverage [%] is shown as a color gradient.

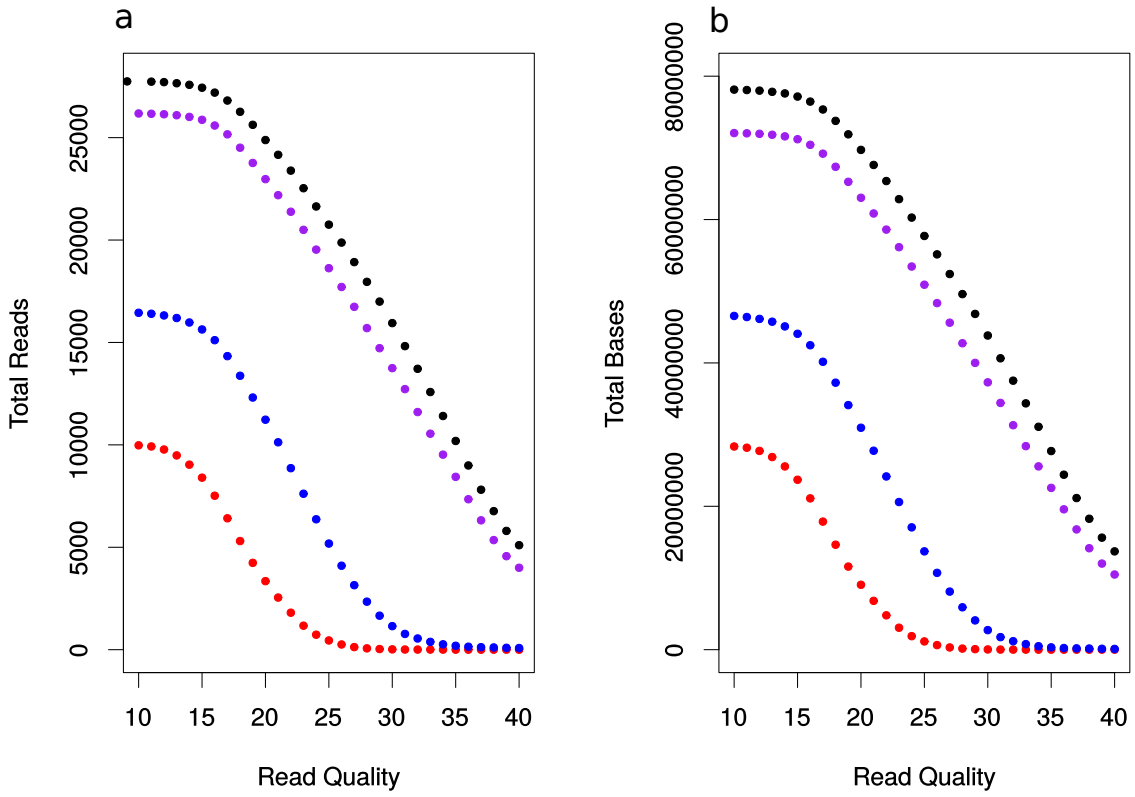


Figure 2.2.: Input statistics
a) Number of sequences ('Total Reads') given different 'Read Quality' cutoffs. For four accessions (Col-0 (black), Cvi-0 (red), KBS-Mac74 (blue), and Ty-1 (purple)), the total number of reads is plotted after filtering using Read Quality cutoffs between 10 and 40. b)Number of bases ('Total Bases') given different 'Read Quality' cutoffs. The total number of bases is plotted for Read Quality cutoffs between 10 and 40.

Quality following a similar slope in all four accessions (fig. 2.2), but Cvi-0 and KBS-Mac74 provided 40 % to 60 % less total input data compared to Col-0 and Ty-1. The mean read length dropped drastically with stricter Quality thresholds for KBS-Mac74 and Cvi-0. In these accessions, long sequences generally had lower Q-scores (fig. 2.3).

For each accession and each read Quality threshold (Q10 - Q40), I assembled using `Canu` (parameters: pacbio-corrected, genomesize=2m, errorRate=0.01, trimReadsCoverage=2) and predicted NLR genes using `AUGUSTUS` (version 3.1.0; –species arabidopsis; Stanke et al. 2004) and `hmmsearch` (`HMMER` 3.1b1; –noali –cut_tc; Eddy (2011)). Whole-genome sequencing based pseudo-chromosomes for the four accessions were created in an ongoing study dedicated to produce accession-specific high quality *A. thaliana* reference genomes. These genomes were used for Quality control in my RenSeq study. They were annotated (`AUGUSTUS`) and NLRs were defined (hmmsearch). The RenSeq assem-

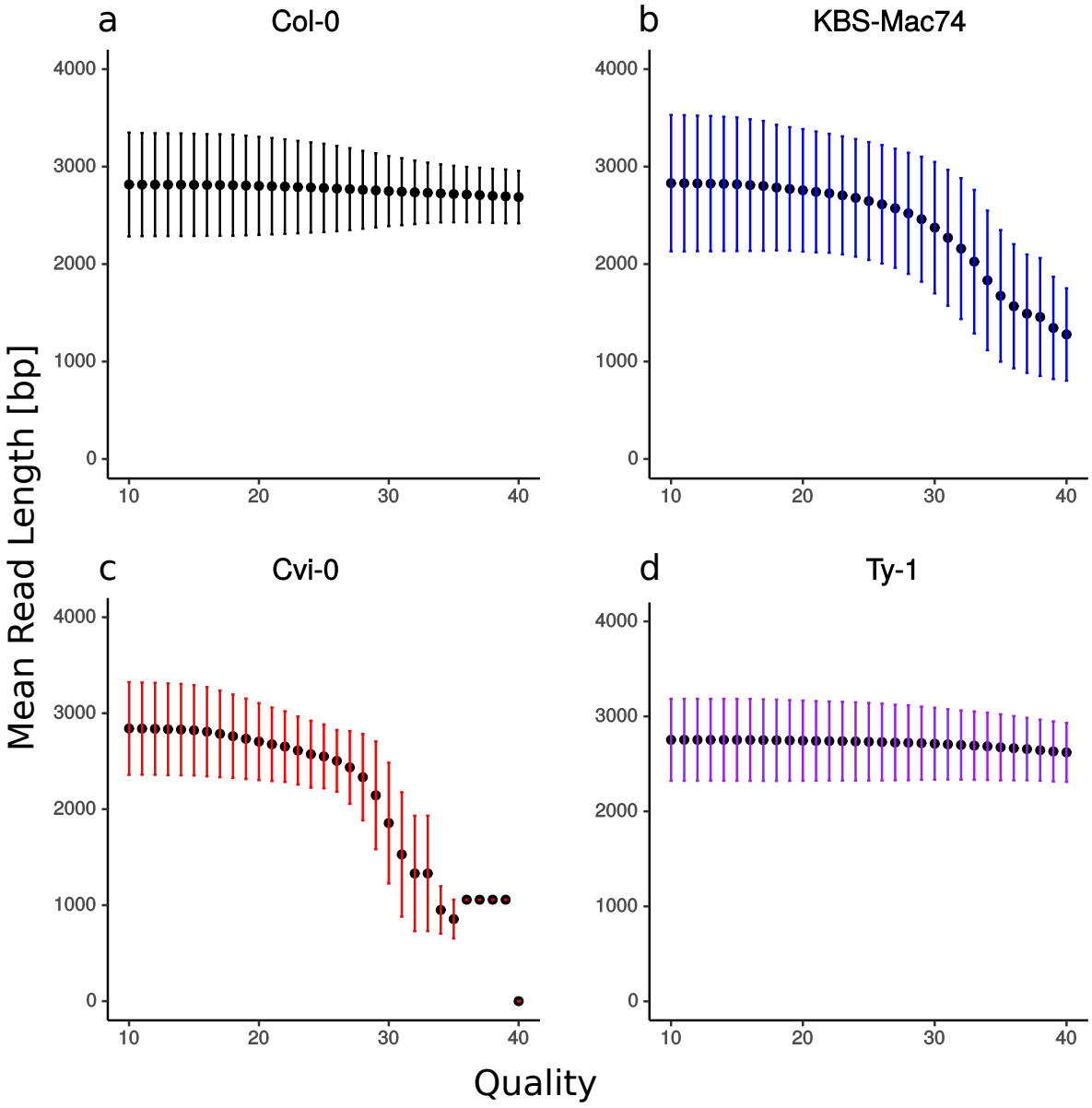


Figure 2.3.: Read length distribution
'Mean Read Length' for different read 'Quality' cutoffs. Mean (black dots) and standard deviation (colored vertical lines) is shown for four different accessions in four subpanels.

blies were compared to the respective pseudo-chromosomes (`Quast`; version 3.2; defaults; Gurevich et al. 2013). I analyzed the assembly size (fig. 2.4), fully and partially assembled NLRs (fig. 2.5), the amount of misassembled contigs (fig. 2.6) and mismatched bases (fig. 2.7).

The RenSeq datasets of Col-0 and Ty-1 provided more than 25 k reads, and more than 15 k of high quality (>Q20), their assembly sizes were in the range of the expected 2 Mb (fig. 2.4). Nearly all NLRs were detected when mapping to the respective reference genomes, but some of them were only partially mapped (>1 base missing, fig. 2.5). Extensive Quality filtering (>Q30) increased the amount of partial NLRs, a result of the shrinking assembly size. Independent from Quality filtering, `Canu` assembled correctly and without many positional errors. Less than 10 contigs were misassembled in Col-0 and Ty-1, and of those, at most two contained an NLR gene (fig. 2.6). With low Quality filtering, around 100 mismatched bases (fig. 2.7) were found in NLR-containing contigs (∼1.2 Mb total sequence), and roughly 500 more in non-NLR containing contigs (less than 0.5 Mb total sequence). Filtering out low quality reads prior to the assembly reduced especially mismatches in non-NLR contigs, but also mismatches in NLR-containing contigs dropped with increased filtering. Some high filtering thresholds (Q=34,35,38,39,40) increased the mismatch rate in Col-0. This increase was not confirmed in Ty-1 and rather reflected random assembly or mapping errors than a general trend.

Cvi-0 and KBS-Mac74 provided fewer reads and more erroneous longer reads (fig. 2.2 and fig. 2.3). Thus their assemblies were smaller and a smaller proportion of NLRs was found. In addition, the found NLRs were more often partial. Still, intermediate filtering (≤Q20) improved assembly size and NLR statistics. With stricter quality filters, `Canu` could not produce assemblies for Cvi-0 (Q≥25) and KBS-Mac74 (Q≥32) because no reads remained. Around 40 misassemblies independent from filtering and up to 10 k mismatches confirmed the bad quality of the KBS-Mac74 assembly. Cvi-0 showed nearly no misassemblies. Given the small total assembly size and the high rate of partial NLRs, this suggested a rather fragmented assembly whose input data did not allow for a high contiguity (see also Cvi-0 (accession identifier 6911) in fig. 2.9). Independent from the filtering, the Cvi-0 assemblies had ∼50 mismatches in NLR contigs and non-NLR contigs. The testing results suggested that good assemblies require many reads (>10 k) of high quality (>Q20) as input for `Canu`. This minimized misassemblies and mismatches, and allowed for the detection of nearly all NLRs. Without a reference genome to validate assemblies and annotations, accessions with few and/or low quality input data would provide incomplete NLR complements, and likely prevent the correct analysis of the pan-NLR'ome. Found NLRs might be correctly assembled and annotated, but the frequency of errors would be higher. Even worse, missing NLRs would in many cases not relate to presence-absence polymorphisms, but rather be artifacts resulting from the bad input data.

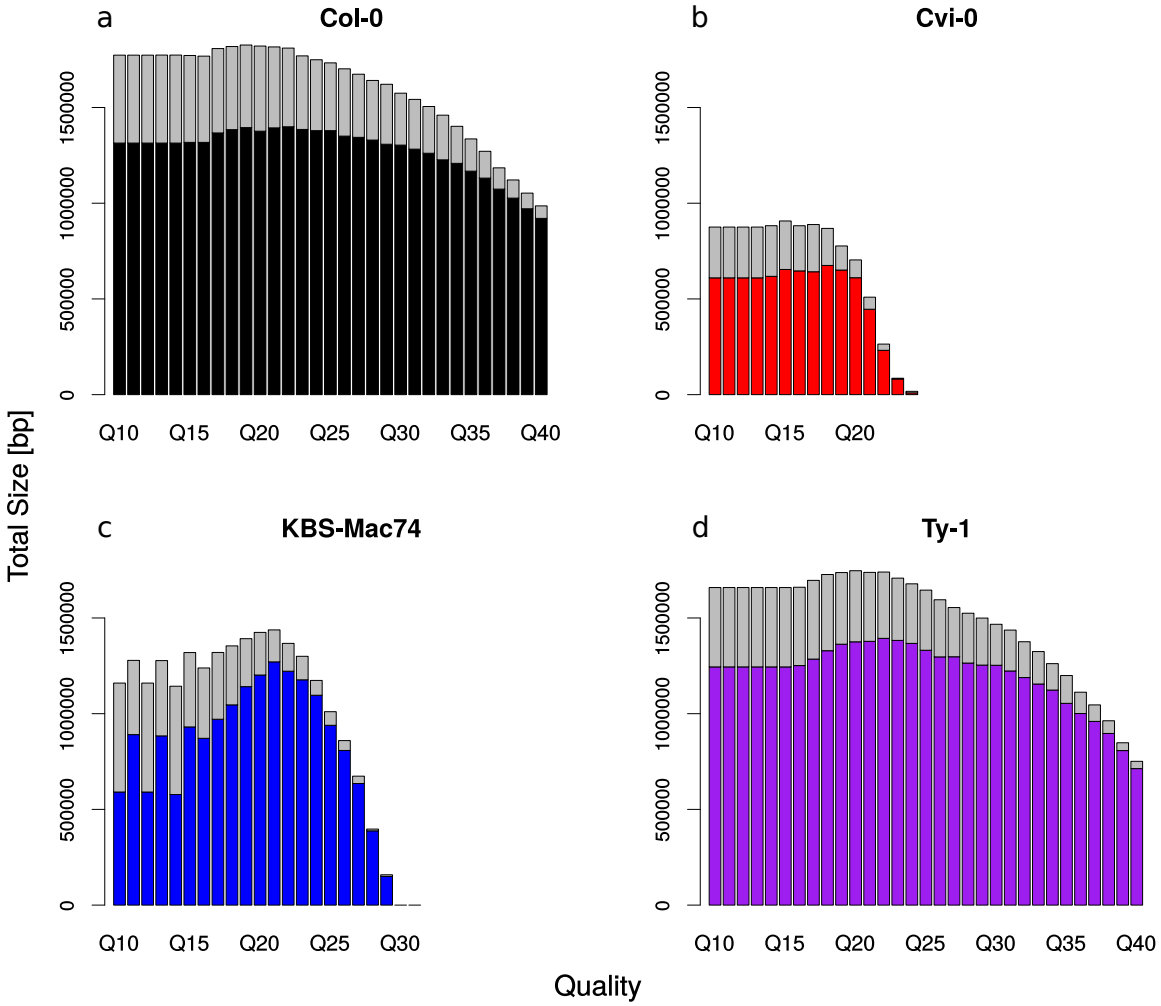


Figure 2.4.: Assembly size
Total assembly size ('Total Size') for different Quality thresholds. For each accession (subpanels) and Quality (x-Axis), NLR-containing contigs (colored bottom part of stacked bars) and non-NLR contigs (grey top part of stacked bars) sum up to the total assembly size.

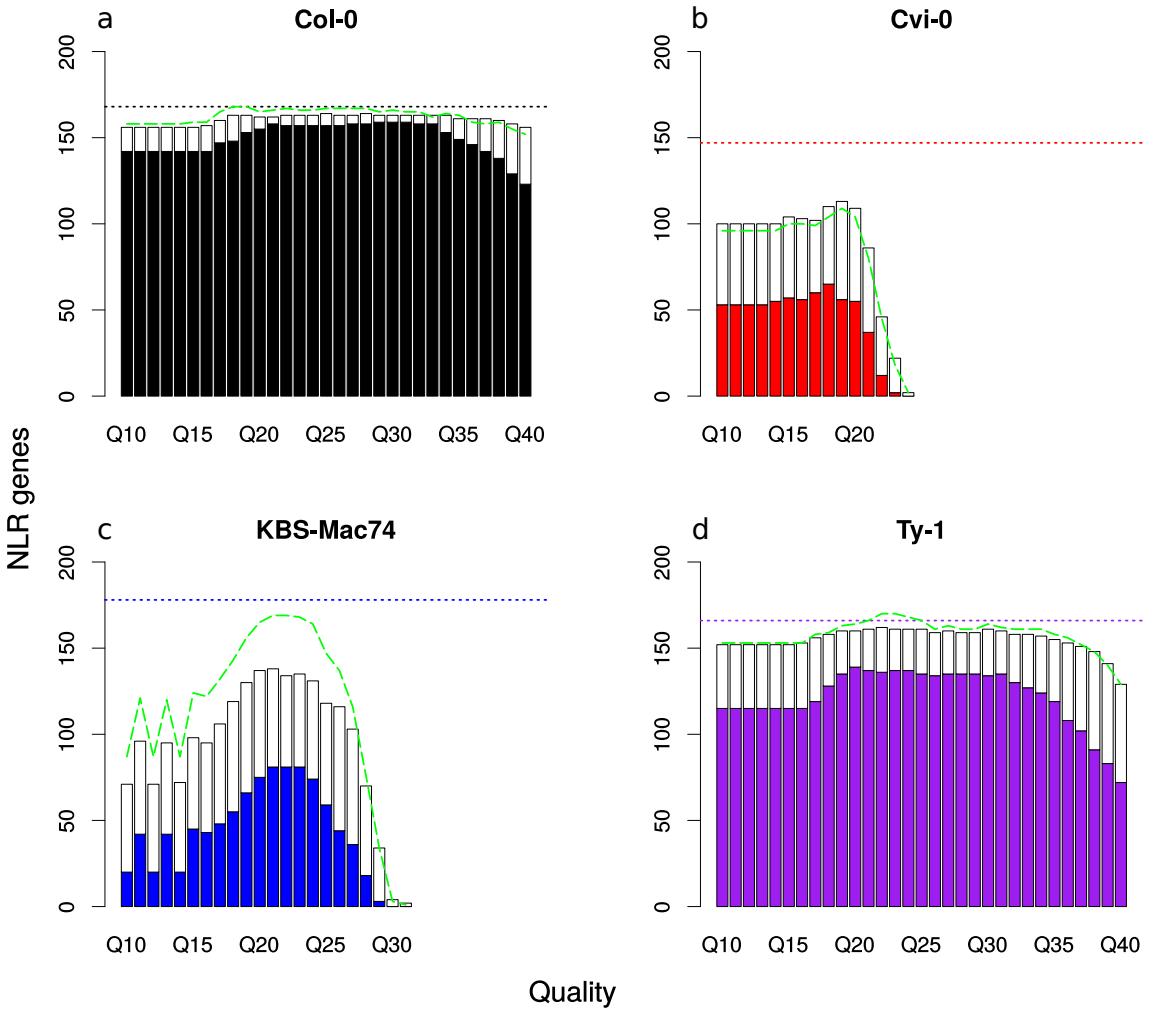


Figure 2.5.: NLR genes
Number of 'NLR genes' for different 'Quality' thresholds. For each accession (subpanels) and Quality (x-Axis), complete NLRs (colored bottom part of stacked bars) and partial NLRs (white top part of stacked bars) sum up to the total mappable NLR gene number. The green dashed line denotes the amount of annotated NLRs (mapped and unmapped) in the RenSeq assemblies, and the dotted line shows the amount of expected NLRs from the pseudochromosomes.

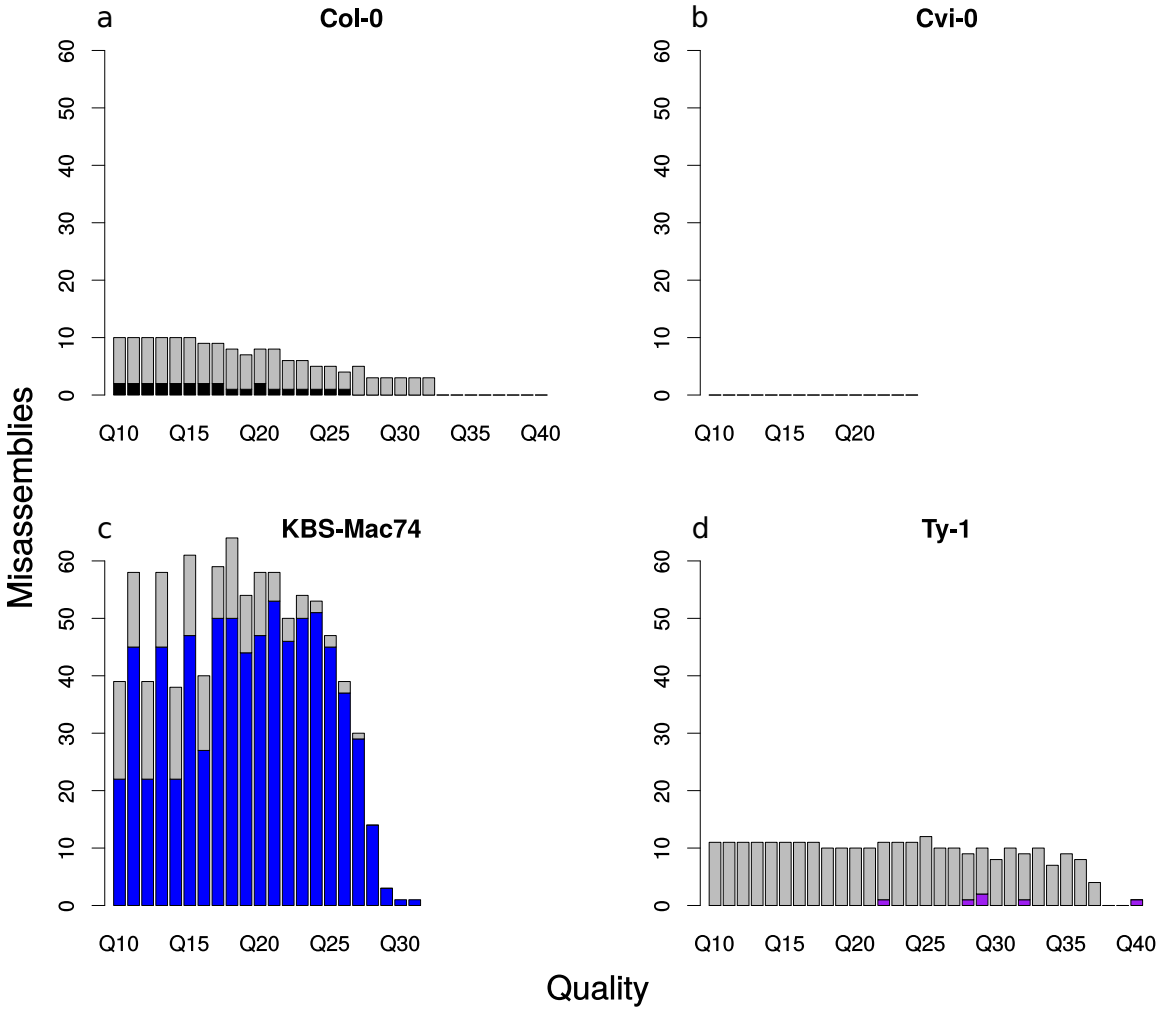


Figure 2.6.: Misassemblies
'Misassemblies' for different 'Quality' thresholds. For each accession (subpanels) and Quality (x-Axis), misassembled contigs with NLRs (colored bottom part of stacked bars) or without NLRs (grey top part of stacked bars) are shown.

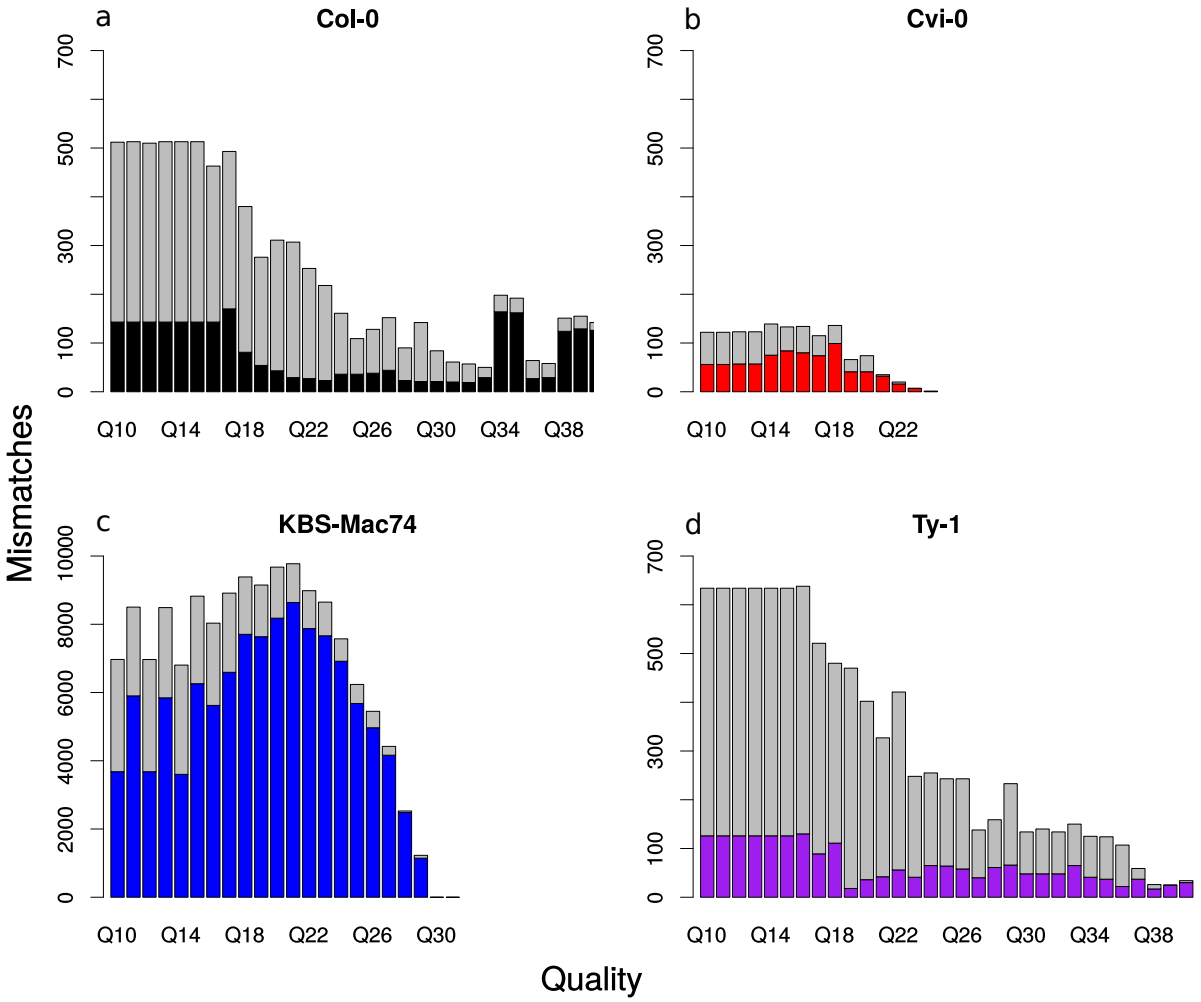


Figure 2.7.: Mismatches
'Mismatches' for different 'Quality' thresholds. For each accession (subpanels) and Quality (x-Axis), the number of mismatched bases are shown in contigs with NLRs (colored bottom part of stacked bars), and in contigs without NLRs (grey top part of stacked bars). Note the different scale for KBS-Mac74.

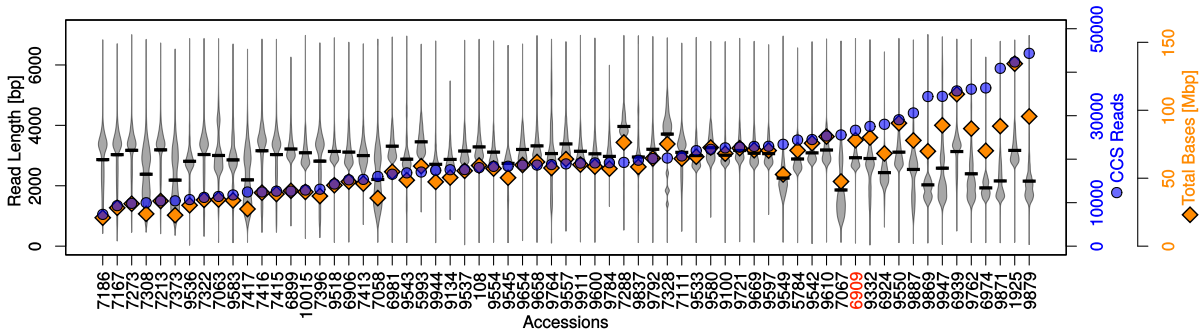


Figure 2.8.: Input reads, read lengths, and total bases for 65 accessions
Read length distribution of corrected and filtered CCS reads (Q20) for the 65 accessions with successful RenSeq experiments. The mean is shown as a solid black horizontal line. The full densities are shown as bean plots (filled grey). The total number of CCS reads (blue circles) and the total number of bases (orange diamonds) are plotted with separate y-axes (see right hand side).The reference accession Col-0 (6909) is highlighted in red.

2.1.3. Assembly of 73 diverse *A. thaliana* accessions

The newly gained knowledge about optimal RenSeq assemblies was applied to a set of 73 *A. thaliana* accessions that were sequenced using SMRT RenSeq in three labs (table A.1 and table A.2).

All but four accessions provided ≥10 k reads with a minimum read quality of Q=20 (at most 1 % error). The sequenced fragments were ∼3kb long (most between 2-5kb). Note that two of the four above tested accessions are not part of this dataset (Ty-1 and KBS-Mac74). I assembled all of them using the above described assembly parameters (see also section 3.7.1.5 and fig. 3.A.13). Five accessions (Cvi-0 (6911), Mar-1 (9555), Vim-0 (9598), Ven-0 (9905), Cat-0 (9832)) were removed from further analyses because of their small assembly sizes (fig. 2.9). Three accessions (Col-0, Ws-2, Can-0) were sequenced in two labs and only the dataset resulting in the better assembly was kept. The remaining 65 accessions represented successful RenSeq experiments and reliable assemblies. For a summary of the sequencing read statistics of those 65 accessions, see fig. 2.8.

I showed in this section why a pipeline for the assembly of RenSeq datasets needs to be conservative and strict. Sensitivity can only be increased if one is willing to put a lot of manual effort into quality control and validation of the NLR gene annotation.

2.2. Annotation Optimization

The primary task after successfully assembling the 65 *A. thaliana* accessions was the gene annotation. The assembled RenSeq contigs contained both NLRs and non-NLR genes. Thus, all genes needed to be predicted and conserved domains needed to be assigned correctly to identify NLR genes for further analyses. Automated gene annotation is inevitable when working with large amounts of sequence data, but the method's accuracy is limited and gets worse when working with incomplete genomes (feature of RenSeq)

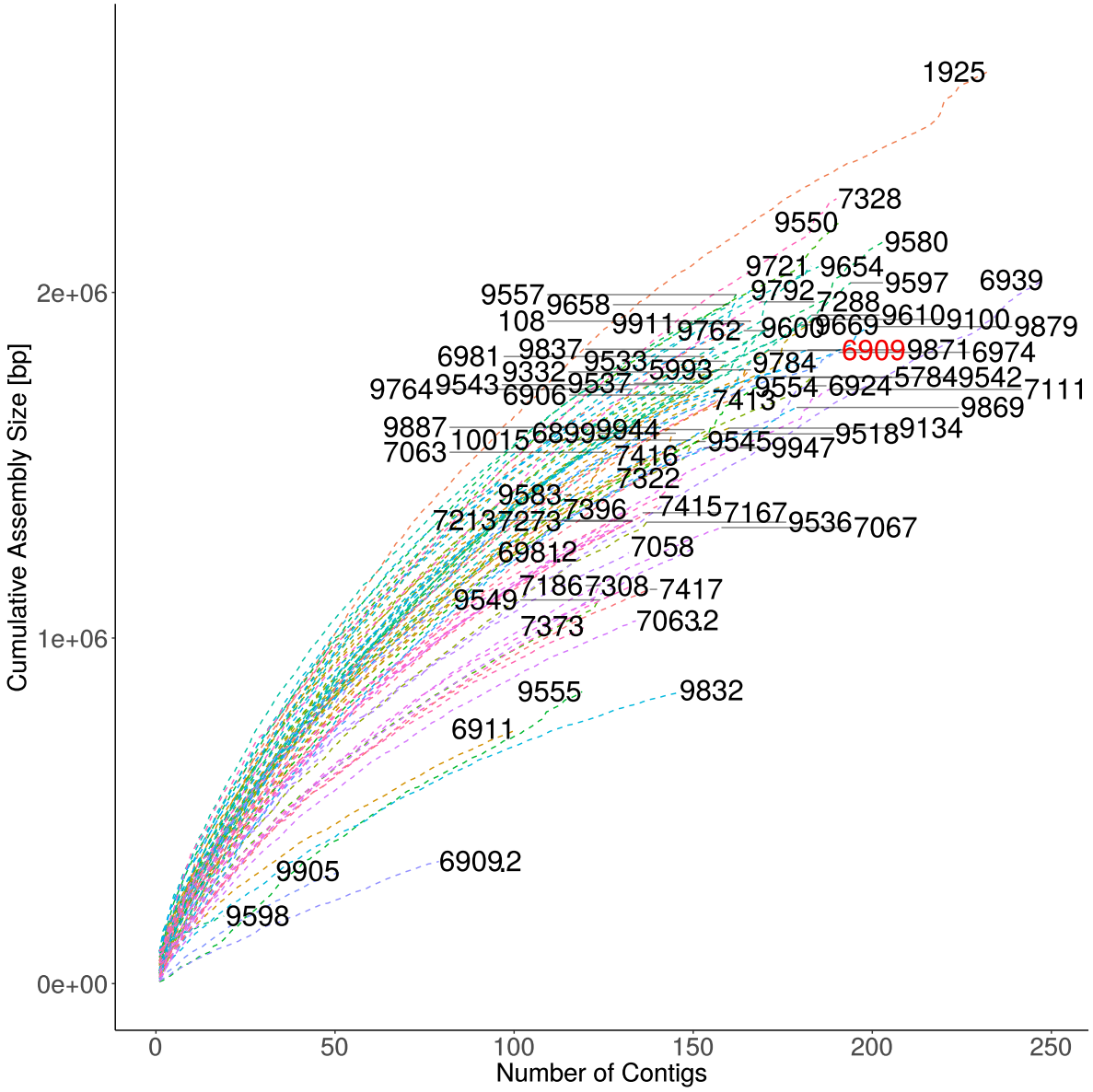


Figure 2.9.: Cumulative assembly size
The 'Cumulative Assembly Size' (y-axis) and the 'Number of Contigs' (x-axis) is shown for 73 RenSeq accessions. Each dashed line corresponds to an individual accession, and the accession identifier is shown at each line's end. Accession identifiers ending with '.2' are used for the three accessions that were removed due to being sequenced in two labs. The RenSeq assembly of the reference accession Col-0 is colored red.

and complex gene families like the NLRs. Manual reannotation was the key to secure a high quality NLR gene complement for *A. thaliana*, and for two outgroups *Arabidopsis lyrata* and *Capsella rubella*.

2.2.1. Automated gene annotation

For automated gene annotation of the 65 accessions, I used the genome annotation pipeline `MAKER` (version 2.3.2, Campbell et al. 2014). It contains a rich set of features including several independent gene predictors, repeat masking prevents unreliable predictions, and known protein and transcript evidences can be incorporated to refine and improve the *ab initio* gene models. Two *ab initio* gene predictors were used. `AUGUSTUS` (version 3.1.0; defaults; Koren et al. (2017)) comes with a special arabidopsis-centered profile that I used for the gene prediction. `SNAP` (version 2006-07-28; defaults, Korf 2004) was trained with a custom hidden markov model (HMM) based on NB and/or TIR domain containing genes from *A. thaliana* to improve the accuracy of NLR gene prediction. The accuracy of exon-intron structures predicted by ab initio methods is reported to be around 60 % to 70 % (Yandell et al. 2012). In a test using the reference Col-0, `AUGUSTUS` sometimes tended towards predictions that were longer than the correct gene and also to fuse genes. `SNAP` showed opposite errors with shortened or split genes. The *ab initio* predictions were thus refined using all annotated Col-0 proteins and transcripts from `Araport11` (Araport11_genes.20151202.pep.fasta, Araport11_genes.20151202.mRNA.fasta; *ARAPORT: Arabidopsis Information Portal* 2018; Cheng et al. 2017). The proteins and transcripts were mapped to the assembled contigs (`MAKER`) with strict settings for the minimum mapping quality (ep_score_limit =95, en_score_limit =95), to secure that only correctly mapped proteins and transcripts were considered as evidence. The specificity of evidence mappings was further increased by restricting the flanking regions around gene models for which evidence was considered (pred_flank=150). Not every gene prediction overlapped with protein or transcript evidence. `MAKER`'s default in those cases was to exclude the prediction. This behavior artificially removes genes that are absent in the reference or genes with pronounced sequence changes, so I decided to keep one of the *ab initio* predictions instead (keep_preds=1). The maximum intron size occurring in the `Araport11` gene set was set in `MAKER` (split_hit=3200). Gene predictions with larger intron sizes were split. Repeat masking improved the annotation accuracy tremendously. Unmasked repeats attract false evidence alignments leading to wrong gene annotations. If not masked prior to gene annotation, transposon open reading frames (ORFs) are often mis-interpreted as exons of neighboring genes, which corrupts the gene annotation further. In a test annotation of the full Col-0 reference genome (27 667 transcripts), masking repeats provided transcript numbers more closely to the reference, resulted in less fragmented annotations (average alignment length is bigger), less wrong annotations (less unannotated transcripts are annotated) and less misannotated transcript positions (average number of mismatches per transcript is lower table 2.2.

In addition to the annotation of the RenSeq assemblies, I also used `MAKER` to revise the NLR gene annotations for the two outgroups *C. rubella* and *A. lyrata* (section 3.7.1.6).

Table 2.2.: Repeat masking
Repeat masking positively influences the gene annotation. Transcript-based statistics of MAKER annotations of the full reference Col-0 genome with or without repeat masking are shown.

	with repeat masking	without repeat masking
Transcripts	29 111	34 838
Avg. alignment length [bp]	1201	1179
Avg. number of mismatches per transcript	0.002	0.01
Unannotated transcripts	2143	7566

2.2.1.1. Limits of automated gene annotation

Automated gene annotation is a standard task for genome-centered research, and current methods provide parameters that can be adjusted depending on the input data and the research question to obtain the best possible result. Still, annotation is not error-free and some NLR- and RenSeq-specific features hamper the task. Gene annotation is commonly performed on whole genomes, but my assemblies were based only on the genomic proportion that contained NLR genes, which meant the contigs were smaller and there might be truncated genes. The NLR-centered RenSeq assemblies also showed the accumulation of a general annotation problem: genes like clustered NLRs, which are close to each other and show a similar structure, tended to get fused. Sometimes, AUGUSTUS or SNAP erroneously predicted an intron instead of detecting the correct gene end, and neighboring genes got fused (fig. 2.10). Fusions also occurred if evidence mappings spanned neighboring genes and thus incorrectly guided MAKER to fuse the genes. The opposite, incorrectly split genes also occurred, mainly for genes without protein or transcript evidence. Pseudogenes did not have protein or transcript evidence, which exacerbated the MAKER-internal decision which gene model to choose. In addition, pseudogenes in general accumulate mutations that deviate from the expected gene composition (e.g. internal stop codons). SNAP and AUGUSTUS tended to fuse pseudogenes to neighboring NLRs instead of predicting the correct pseudogene (fig. 2.11). Seldom, misassembled contigs lead to misannotated genes, and even rarer, erroneous reference annotations misguided MAKER's decision making process. Taken together, the here mentioned problems lead to a significant number of misannotated genes that had to be corrected manually.

2.2.2. Manual gene reannotation

I developed, together with collaboration partners (chapter 3), an SOP for manual reannotation (section 3.C.1) using the genome annotation editor WebApollo (version 2.0.4; http://ann-nblrrome.tuebingen.mpg.de/annotator/index; Lee et al. 2013), and reannotated all gene models that contained an NB or TIR domain in the RenSeq datasets and in the outgroups. The final MAKER gene models were integrated into Web Apollo, as well

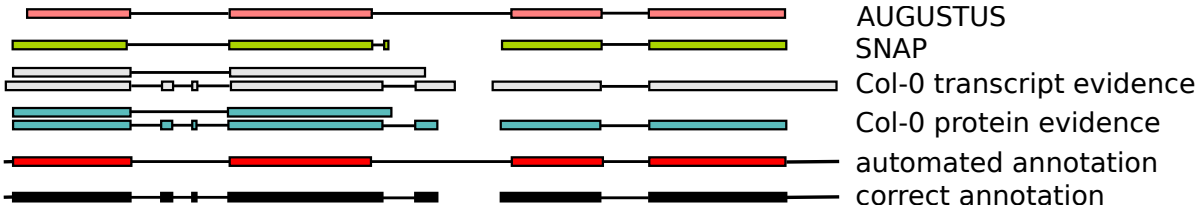


Figure 2.10.: NLR fusion detected with Col-0 protein and transcript evidence
Scheme of an NLR fusion. Shown are the `Web Apollo` tracks for the final automated `MAKER` annotation (red), the two gene predictors (`AUGUSTUS` lachs and `SNAP` green), Col-0 protein evidence (blue), Col-0 transcript evidence (grey), and the true (but unknown) gene structure (black).

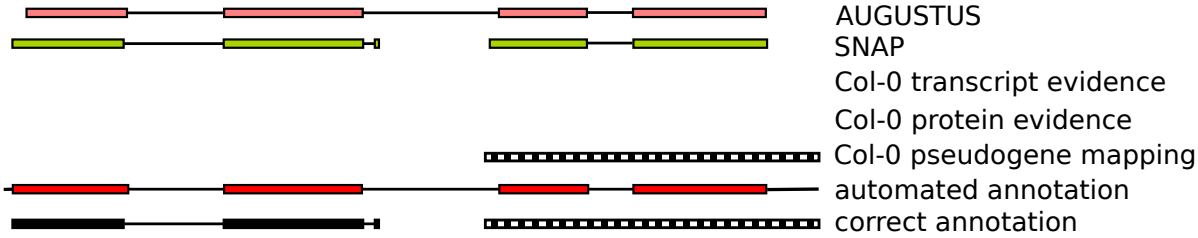


Figure 2.11.: NLR fusion detected with a pseudogene mapping
Scheme of an NLR fusion. Shown are the `Web Apollo` tracks for he final automated `MAKER` annotation (red), the two gene predictors (`AUGUSTUS` lachs and `SNAP` green), pseudogene mappings (black-white), and the true (but unknown) gene structure (black).

as all underlying predictions (`AUGUSTUS` and `SNAP` predicted genes) and evidence lines (proteins and transcripts with a minimum mapping score 95 %). Relaxed mapping of the Col-0 evidence proteins, transcripts, and of Col-0 pseudogenes (Exonerate; version 2.2.0; minimum mapping score 50 %; Slater et al. 2005;) supported the reannotation of duplicated and diversified genes. Protein domains guided the NLR reannotation further. They were predicted both for the final `MAKER` gene models, and for the `AUGUSTUS` gene predictions using `Pfam` HMMs and coiled coils (`InterProScan`; version 5.20-59.0; -dp -iprlookup -appl `Pfam`,`Coils`; Zdobnov et al. 2001). Repeats that were masked by `MAKER` to facilitate the gene annotation were included in `Web Apollo`, too. The general reliability of a contig was reflected using raw read support. Positions were more reliable if many CCS reads could be mapped (`pbalign`; version 3.0; defaults; Tyagi et al. 2008) and if many of those reads were used by `Canu` to assemble the position. Mappings and reads used for the assembly were shown in `Web Apollo`. Blast alignments for the databases nr (Non-redundant protein sequences) and nt (Nucleotide collection) also helped guiding the reannotation.

All 13 911 gene models containing NB-ARC or TIR domains were manually inspected. Errors were corrected in 4199 of those genes. Gene fusions were inferred using Col-0 protein and transcript mappings. Genes were split if two evidences (protein, transcript, or pseudogene) mapped next to each other within one gene model (fig. 2.10). Further sup-

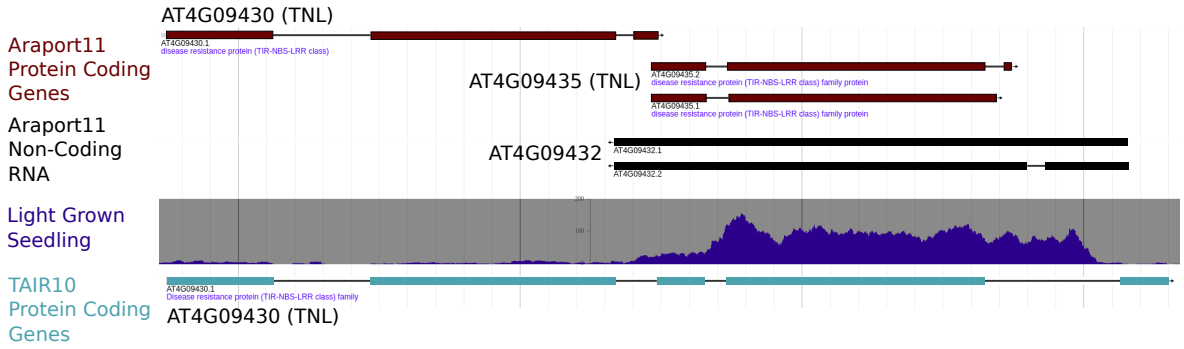


Figure 2.12.: Misannotated reference gene *AT4G09430*
Adapted screen shot from the `Araport11` Annotation Release (*Araport11 JBrowse* 2018) The genes *AT4G09430* and *AT4G09435* are incorrectly annotated in `Araport11`. The original `TAIR10` annotation shows the correct annotation and should not have been split. Expression levels are shown for RNA-Seq read mappings from Light Grown Seedlings of Col-0 (*Araport11 JBrowse* 2018; Cheng et al. 2017).

port for a gene fusion often was given by an unusually long intron spanning the actual intergenic region. 736 NLR genes were found erroneously fused and were corrected to 1473 manually split NLRs. Incorrectly split genes were also detected using Col-0 proteins and transcripts. Genes were merged if one Col-0 protein or transcript mapping traversed several neighboring gene models. In total 247 genes had to be merged. Truncated genes confirmed by protein or transcript mapping, were found in 453 cases. They were flagged and gene models were extended towards contig borders where possible. Gene-, exon- or intron-boundaries were refined in numerous cases using protein, transcript, and RNA-seq mappings. Also without direct mapping evidence, boundaries were sometimes changed in order to obtain the open reading frame containing a domain annotation (97 cases). These reannotations had to be confirmed by at least one other researcher and they were flagged ('corbound'=changed exon-/intron boundaries; 'cortrans'=changed translation start/end) to be able to distinguish them from evidence-based reannotations. Noncanonical splice sites had to be introduced manually if reference protein or transcript evidence existed, because `MAKER` only models canonical splice sites. Rarely, erroneous reference annotations were detected leading to incorrect gene models. As an example, *AT4G16857* is annotated as a TNL, but does not contain any domains. *AT4G09430* is wrongly split into two TNLs (*AT4G09430* and *AT4G09435*, fig. 2.12). The first gene contains the NB- and the TIR- domain, the second one only has an LRR annotated. The misannotation is driven by a natural antisense gene (*AT4G09432*) that overlaps *AT4G09430*. This antisense gene is expressed, whereas *AT4G09430* might not be expressed (at least not in the data used for `Araport11` annotation creation). This expression is mistreated as belonging to the TNL, which results in the split. In cases of misannotated reference NLRs, genes were corrected using the TAIR10 annotations (Berardini et al. 2015). In addition to the reannotation of erroneous gene models, manual curation allowed for the detection of gene-centered features of the RenSeq datasets. Neighboring NLRs in head-to-head

orientation were flagged 'paired' because literature suggests co-operative function for some of them (Narusaka et al. 2009). Independent of the orientation, neighboring NLRs with $> 95\,\%$ similarity to a known NLR pair from `Araport11` were also flagged. In total 1655 genes were found paired in the RenSeq assemblies. Putative pseudogenes were flagged using mapping evidence from known pseudogenes in `Araport11`. Assembly errors were detected and affected genes were flagged if CCS read mappings reliably proved a misassembled position.

I combined optimized automated gene annotation with extensive manual curation of NLR genes. This resulted in highly reliable NLR gene complements for all RenSeq accessions, which were necessary to describe the *A. thaliana* pan-NLR'ome.

2.3. NLR Classification: coiled-coils (CCs)

NLRs are typically classified based on their conserved domains. Protein domains were predicted using `Pfam` HMMs as mentioned earlier (section 2.2). I refined coiled-coil (CC) motifs in NLR genes using a majority vote validation with three different programs to account for the inaccuracy of the individual methods. `Coils` (2.2.1; `InterProScan`-defaults; Lupas et al. 1991) and `Paircoil2` (defaults; McDonnell et al. 2006) predict CCs using databases of many known coiled-coils, and the `NLR-parser` (v.2; defaults; Steuernagel et al. 2015) uses two NLR-specific coiled-coil motifs (motif16 and motif17).

Figure 2.13 compares the intersections of CC-containing NLR genes predicted by the three programs, and the reference annotation `Araport11`. Of the 25 CC-containing `Araport11` genes, 12 are found in all predictors (rightmost vertical bar), and 10 others are found using two of the three predictors. All programs predict more CC-containing NLRs than are reported in `Araport11`, and there are 20 NLR genes that have a CC motif in at least two of the three prediction methods, but no equivalent in the reference annotation. These results suggest that no method alone is suited to reflect the reference, but it also shows that the reference annotation might not be complete. In addition, I validated how CCs of known functional CNLs are represented by the three predictors, table 2.3 summarizes the results. CCs are often found by all (6 NLRs), or at least two (4 NLRs) of the three methods. Three NLRs are only confirmed by the `NLR-parser`, and *ADR1*, an RNL containing an unusual 'CC_R' domain (Collier et al. 2011), is not found by any method. This suggests again that an overlap of two methods provides a reliable CC prediction while not losing too much sensitivity. I reannotated coiled-coils in all RenSeq NLR genes and trusted the annotation, if overlapping predictions existed for at least two of the three methods. Borders were defined by `Coils` (most sensitive method) or by `Paircoil2` if no `Coils` prediction was present. The `NLR-parser` was not used to define boundaries because it reports only the positions of two motifs known from CCs instead of the full CC. NLRs were then classified into TNLs, CNLs, RNLs, and NLs (section 3.7.1.10, and section 3.7.1.11).

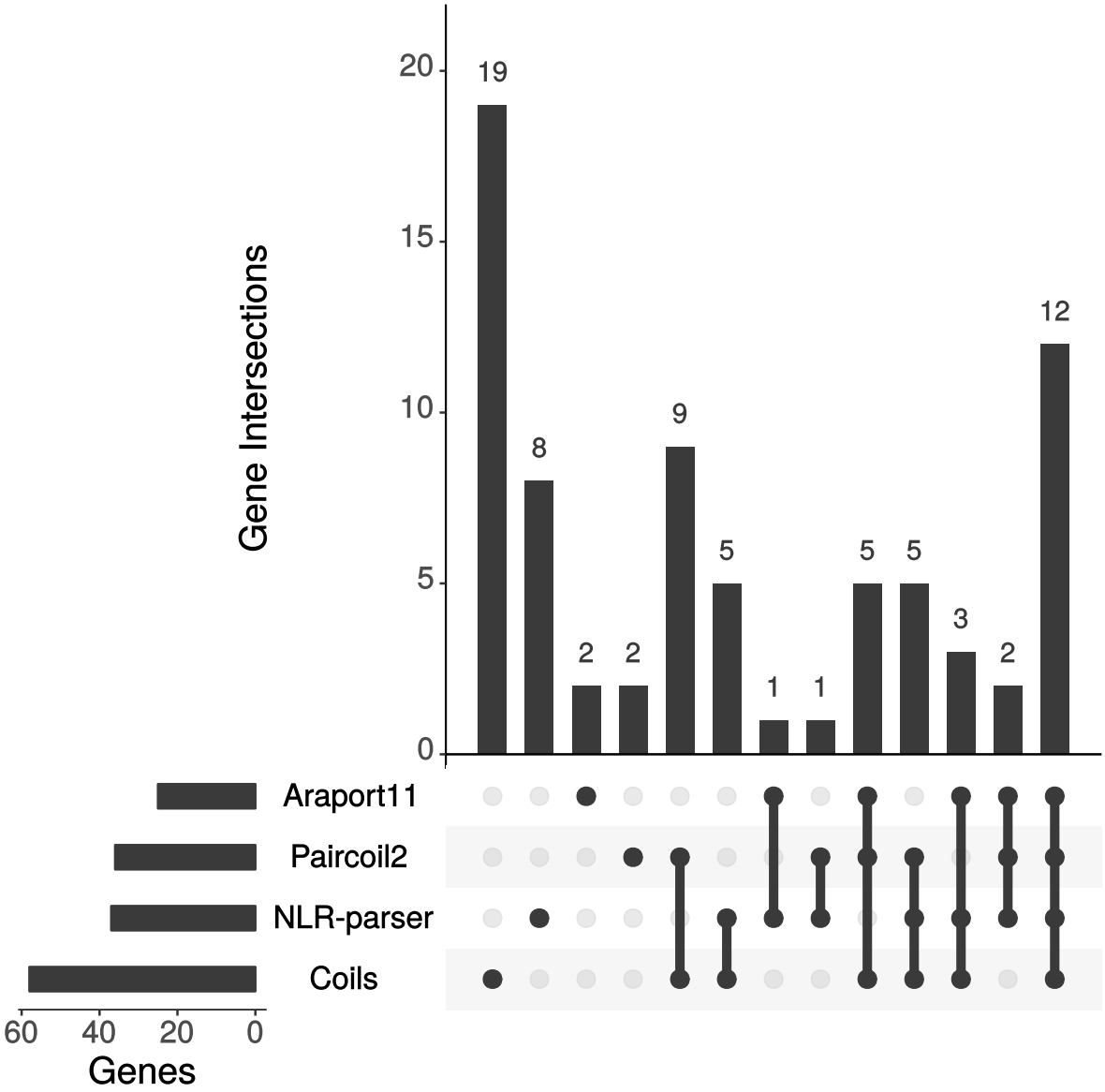


Figure 2.13.: Coiled-coil containing NLRs
Intersection of Col-0 CC-containing NLRs predicted by `Coils`, `Paircoil2`, and the `NLR-parser`, and the reference annotation `Araport11`. The total number of CC-containing NLRs is shown as horizontal bars in the bottom left part of the figure. All intersections are shown in a matrix-layout combined with vertical bars. For each combination (black dots and their connections by lines), the bar shows the number of intersected CC-containing genes.

Table 2.3.: CC detection in known functional CNLs
CC predictions (binary, 0/1) from `Paircoil2`, `Coils`, and the `NLR-parser` for known functional CNLs. The `Araport11` identifier and the name of each CNL is given.

Identifier	Name	`Paircoil2`	`Coils`	`NLR-parser`
AT3G07040	*RPM1/RPS3*	0	0	1
AT4G26090	*RPS2*	1	0	1
AT1G10920	*LOV1*	0	0	1
AT1G58602	*RPP7*	1	0	1
AT1G12220	*RPS5*	0	1	1
AT5G43470	*HRT/RPP8*	1	1	1
AT1G33560	*ADR1*	0	0	0
AT1G12280	*summ2*	1	1	1
AT1G12210	*RFL1*	1	1	1
AT1G59620	*CW9*	0	0	1
AT1G61180	*Uni-1d (Ws)*	1	1	1
AT1G61190	*RPP39*	1	1	1
AT3G46530	*RPP13*	0	1	1
AT3G50950	*ZAR1*	1	1	1

2.4. Assembly Quality and NLR complement Completeness

I evaluated the quality of an assembly by comparing how good the assembly fit its input reads. I assessed the quality using pseudo-heterozygous SNP calls that were created by mismapped CCS reads. If an NLR was not correctly assembled, CCS reads from that gene mapped to a similar NLR instead and created pseudo-heterozygous SNPs. A 'Quality' value was calculated from the ratio of those SNPs to the total amount of mapped NLR gene bases. For each RenSeq assembly, I created a pseudo-genome consisting of all assembled contigs and the `TAIR10` reference chromosomes. All NLRs were masked on the Col-0 reference chromosomes to avoid reference-biased NLR gene mappings. All non-NLRs were masked on the RenSeq contigs to secure only one representation of all non-NLR genes existed in the pseudo-genome (fig. 2.14). I then mapped the accession's CCS reads to the pseudo-genome (`Minimap2`; 2.9-r748-dirty; -x map-pb; Li 2018b) and called SNPs for NLR genes using high quality mappings only (`htsbox` pileup; r345; -S250 -q20 -Q3 -s5; Li (n.d.)). The Quality was calculated using a formula adapted from the Phred-quality score, which is used to determine per base qualities (eq. (2.1)). The Quality (Q) is logarithmically linked to the ratio of pseudo-heterozygous calls (hetsites) and the total amount of mapped NLR gene bases (totalsites).

$$Q = abs(-10 * log10(\frac{hetsites}{totalsites})) \tag{2.1}$$

Figure 2.14.: Diagram for pseudo-genome generation
Pseudo-genomes contain the assembled RenSeq contigs with masked (purple) non-NLR genes and unmasked (blue) NLRs. In addition, the `TAIR10` reference chromosomes are used, with masked (purple) NLRs and unmasked (blue) non-NLRs.

Quality values of RenSeq assemblies ranged between 17.7 and 42.5 (median=29.7), with Col-0 showing an intermediate Quality of Q=30.6 (fig. 2.16 panel b, black dots). I checked whether the Quality was correlated with the input reads, the total number of bases or the read length N50, which would suggest that accessions with lower Quality were limited by those input features. I also checked if an accession's Quality was correlated with the similarity to Col-0 (see section 3.7.5.3 for identity by state (IBS) calculation method), which would suggest a bias in the efficiency of the baits that were mostly created from Col-0 NLRs. The sub-sampling test showed erratic high Qualities for small sub-sets (fig. 2.16, panel a). To avoid using erratic high Qualities of small RenSeq assemblies, I excluded accessions with less than 11.8 Mbp input and a Quality Q<20 (corresponds to 15 % sub-sampling). No positive correlation was detected between the Quality, the input reads, the total number of input bases or the read length N50 (fig. 2.15). This confirmed that enough input fragments were sequenced to secure a reliable assembly. There was also no positive correlation between the Quality and the similarity to Col-0, which confirms the unbiased enrichment capability of the used baits.

Quality values could only be compared to each other and allowed statements about the relative arrangement of accessions. I needed in addition an absolute value of the 'Completeness' of each accession's NLR complement. Without reference genomes, an accession's Completeness could only be derived from the reference Col-0. I used the correlation between Quality and Completeness of Col-0 sub-assemblies to infer the Completeness of RenSeq assemblies. I sub-sampled the Col-0 CCS reads from 100 % (26 639 reads, N50=2846 bp, 77.98 Mb total) to 1 % in 1 % steps (`seqtk` sample; v.1.0-r82-dirty; defaults; Li (2018a)). I assembled the sub-sampled datasets (section 2.1). NLRs were annotated by mapping all NLRs from the full RenSeq Col-0 annotation to each sub-assembly. NLR transcripts were extracted (`exonerate`; v.2.2.0; –model est2genome –bestn 1 –refine region –maxintron 546; Slater et al. 2005) and the Quality of each

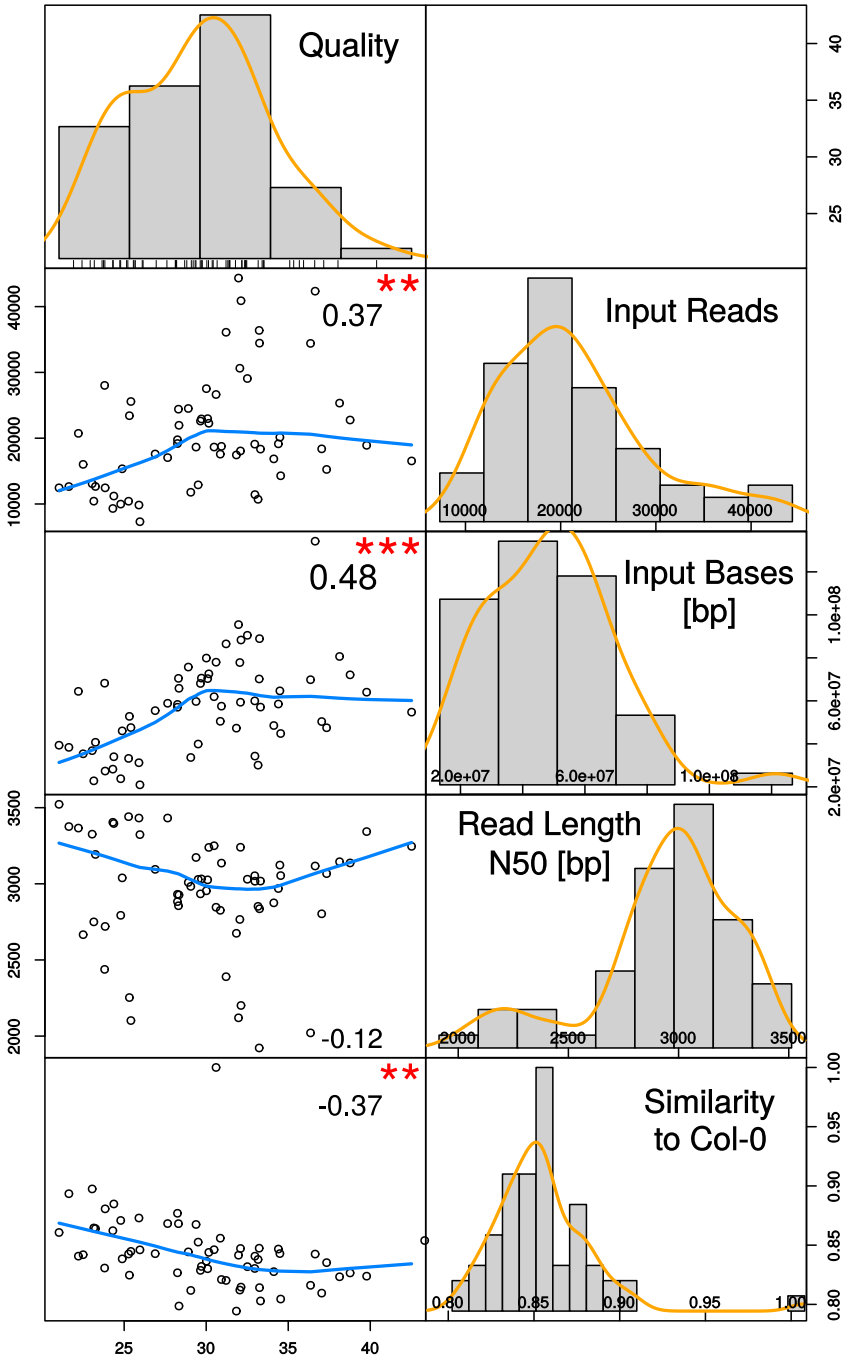


Figure 2.15.: Assembly Quality correlations
Correlations (scatter plots) between the Assembly 'Quality' (top left box), the amount of 'Input Reads', the amount of 'Input Bases [bp]', the 'Read Length N50 [bp]', and the 'Similarity to Col-0' (boxes on the right) for RenSeq datasets. Histograms and kernel densities (orange lines) are plotted for each variable. Scatter plots for variable pairs are shown together with a fitted line (blue) and the Pearson's correlation coefficient (significance niveau 0 '***',0.001 '**', 0.01 '*', 0.05 '.', 1 ' ').

sub-assembly was assessed as described above. The sub-sampling experiment confirmed increasing assembly Qualities with increasing amount of input data (9 %-100 %, fig. 2.16, panel a)). I excluded the 1 %-8 % sub-sampling experiments: no assembly could be produced using 1 % of the input data. The Quality calculation only makes sense for reasonable amounts of mapped CCS reads, therefore the 2 %-8 % sub-assemblies were deemed unreliable and were not used for the Completeness inference of RenSeq assemblies (see below). The Completeness of a sub-assembly was calculated from the fraction of the reference Col-0 NLR complement (from `Araport11`) that got assembled. For each sub-assembly, NLR transcripts were mapped to the reference (`rnaQUAST`; version 1.5.0; defaults with `TAIR10` reference genome and `Araport11` NLRs; Bushmanova et al. 2016) and the Completeness was calculated by dividing the amount of covered NLR genes [bp] by the total length of the `Araport11` NLRs [bp]. In the sub-sampling experiments I could show a quick Completeness saturation reaching a maximum of 97 % (fig. 2.16 panel a, green dots). Already when using only 29 % of the Col-0 input data, 95 % of the NLR gene complement got assembled. Using 29 % of the input data means 7713 reads with an N50 read length of 2849 (50 % of the data can be found in reads of this length or longer), and an N90 of 2592. This is roughly what the 'worst' RenSeq accession provides as input (fig. 2.13), already hinting at a good overall Quality and Completeness of the RenSeq datasets.

I used the relationship between the Quality and the Completeness of the Col-0 sub-assemblies to infer the Completeness of the RenSeq accessions. The smallest Col-0 sub-assembly with a Quality higher than each RenSeq dataset was defined. I assigned the Completeness of the next smaller sub-sampled assembly to the RenSeq accession.

The Completeness of all 65 RenSeq datasets was generally high (median=95 %), 46 accessions were at least 95 % complete (vertical black line in fig. 2.16 panel b), and 56 accessions were at least 90 % complete. Completeness values ranged between 63 % and 97 %, which reflected the maximum obtainable score equivalent to the full Col-0 dataset. Since the Col-0 dataset did not provide a perfect assembly, 23 accessions with a higher Quality could not be ranked. Their Completeness (unfilled green circles in fig. 2.16 panel b) was between 97 % and 100 %. The Quality and the Completeness analyses of the Col-0 sub-sampling experiment and the RenSeq datasets confirmed the high reliability of the RenSeq method (no bait bias, enough input data per accession) and suggested near-complete NLR complements for many accessions.

2.5. Complete NLR complements for the analysis of the pan-NLR'ome

The extensive method optimization and quality control presented in this chapter resulted in near-complete NLR complements for 65 *A. thaliana* accessions. RenSeq reliably enriched the NLR containing proportion of the genome, the assembly resulted in a contiguous representation of full NLR genes and their surrounding regions, and the combination of automated and manual annotation of NLR genes and domains secured

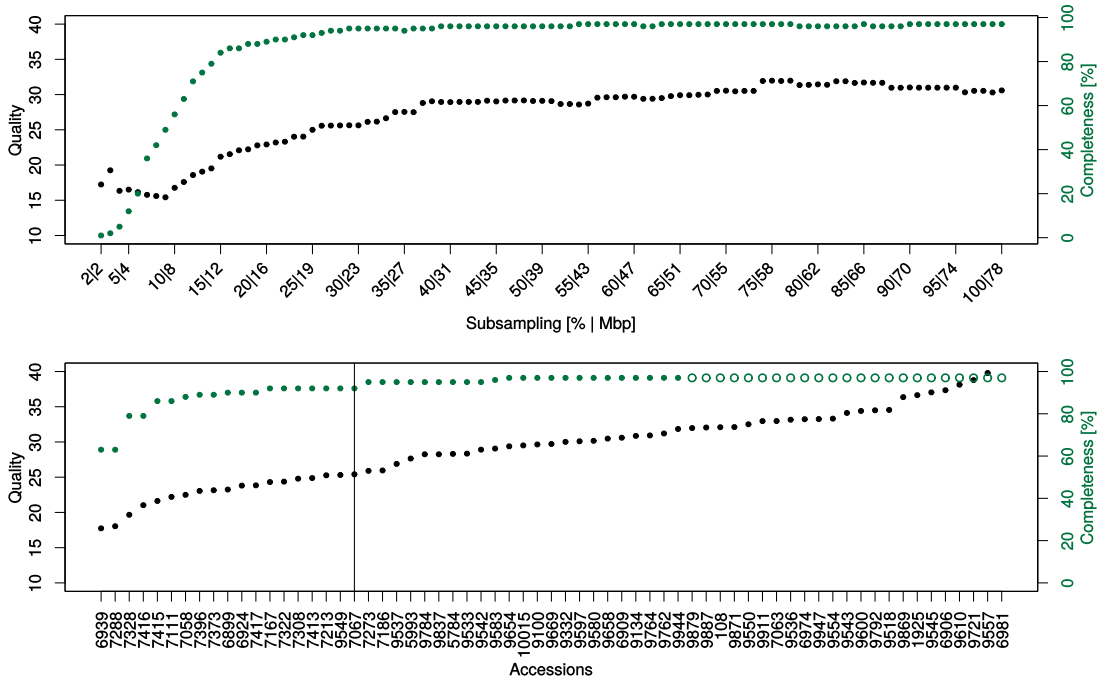


Figure 2.16.: Assembly Quality and Completeness
'Quality' and 'Completeness' of sub-sampling Col-0 test assemblies and RenSeq assemblies. a) Quality (black) and Completeness values (green) for sub-sampled Col-0 datasets. The x-axis contains the sub-sampling percentage and the amount of input data for each sub-sampling experiment. b) Quality (black) and Completeness values (green) for all RenSeq accessions. Unfilled circles indicate accessions with qualities larger than any sub-sampled dataset. The vertical black line is drawn at 95 % Completeness.

correct NLR complements for each accession.

These datasets allowed for an unprecedented and comprehensive description and analysis of the *A. thaliana* pan-NLR'ome, presented in the following chapter containing the manuscript The *Arabidopsis thaliana* pan-NLR'ome. The biological analyses included the architectural diversity of NLRs, especially the use of novel IDs and novel NLR domain architectures. Further, the pan-NLR'ome was shown to be saturated, allowing for a definition of the core NLR complement, and the description of presence-absence polymorphisms in non-core NLRs. Haplotype saturation could be shown, too. The selective forces acting on NLRs, domains, or positions, were quantified and evolutionary coupled co-evolving NLRs could be identified.

3. The *Arabidopsis thaliana* pan-NLR'ome

3.1. Declaration of Contributions

This chapter contains the manuscript The *Arabidopsis thaliana* pan-NLR'ome, which is work in progress and has not been submitted to a journal, yet. Footnotes were introduced to provide further important information to the reader of this thesis. They do not belong to the manuscript.

All authors contributed equally to the scientific ideas presented in the following manuscript. The project was managed mainly by Felix Bemm (FB, last author) and Detlef Weigel (DW, corresponding author), with additional help equally of Freddy Monteiro (FM, shared first author) and Anna-Lena Van de Weyer (AVDW, shared first author). RenSeq data were generated equally by Oliver Furzer (OF, shared first author), FM, Marc Nishimura (MN), and AVDW. Assemblies were produced equally by FB and AVDW. Genes were annotated equally by FB and AVDW, and were manually curated equally by OF, FM, and AVDW, with additional help of FB. Biological analyses were carried out equally by FB, OF, FM, and AVDW. Results were interpreted equally by FB, Jeffery L. Dangl (JD), OF, FM, AVDW, and DW. The initial draft of the paper was written equally by FB, OF, FM, and AVDW. Critical revisions of the manuscript involved FB, JD, OF, Jonathan Jones (JJ), FM, AVDW, and DW.

3.2. Abstract

Plant disease resistance to pathogens is a genetically encoded trait of value for agriculture. Disease resistance is often encoded by nucleotide-binding leucine-rich (NLR)-encoding genes that occur with dynamic incidence across populations. Plant genome sequencing projects provided glimpses of NLR diversity across species, the reported reference gene repertoires are neither exhaustive, nor represent species-wide NLR diversity. Here, we aimed to define the *Arabidopsis thaliana* species-wide NLR'ome. We applied NLR-gene enrichment and long-read sequencing to a collection of 65 diverse accessions, finding that the pan-NLR'ome saturates at approximately 40 accessions. We show that half of the pan-NLR'ome is present in most accessions, whereas the rest exhibited greater variation. We also provide a genome browser and interactive display of phylogenetic relationships within each defined orthogroup. We charted the architectural diversity of NLR proteins and identified novel architectures and integrated domains. We show haplotype saturation and quantify the selective forces that act on specific NLRs, domains,

and positions, contrasting these data with functional annotations. Our results add to the wealth of *A. thaliana* data, while resolving previously uncertain loci. This study defines the concept of the pan-NLR'ome in plants and will be applicable to other plant species.

3.3. Introduction

Plant immune receptor repertoires have been shaped by millenia of plant-microbe coevolution (Gao et al. 2018; Jones et al. 2006). Immunity is activated either by cell surface receptors that recognize microbe-associated molecular patterns (PAMPs), or by intracellular receptors that detect pathogen effectors (Jones et al. 2006). These intracellular receptors are typically encoded by highly polymorphic disease resistance genes. About two thirds of disease resistance genes encode nucleotide-binding leucine-rich repeat receptors (NLRs) (Kourelis et al. 2018), and most plant genomes have hundreds of NLR genes (Shao et al. 2016). The majority of plant NLRs contain a central nucleotide binding domain shared between Apaf-1, Resistance proteins and CED4 (NB-ARC, hereafter NB for simplicity) (Van der Biezen et al. 1998). Most contain also leucine-rich repeats (LRRs) (Maekawa et al. 2011; Takken et al. 2012), and either a Toll/Interleukin-1 receptor (TIR) or coiled-coil (CC) domain at the N-terminus (Bernoux et al. 2011; Nishimura et al. 2017; Qi et al. 2012). Proteins with similar arrangements of functional domains are also involved in host defense in animals and fungi (Jones et al. 2016; Li et al. 2015; Uehling et al. 2017).

Recognition by NLRs generally involves one of three main mechanisms (Dangl et al. 2013). NLRs can directly detect pathogen effectors through interaction with the canonical NLR domains (Catanzariti et al. 2010; Dodds et al. 2006; Krasileva et al. 2010), or with an NLR-incorporated integrated domain (ID) that resembles known domains of pathogen effector targets (Cesari et al. 2014; Le Roux et al. 2015; Maqbool et al. 2015; Sarris et al. 2015; Wu et al. 2015). Alternatively, NLRs detect effector activity indirectly by monitoring a host virulence target ("guardee") (Mackey et al. 2002; Qi et al. 2014; Wang et al. 2015), or detect effectors that interact. Importantly, these mechanisms have been directly demonstrated only for a very small number of NLRs, and additional mechanisms might await discovery.

To date, NLR complements, or NLR'omes, have been defined from available genome annotations for single cultivars of plants or for multiple species across different taxonomic levels, respectively (Bailey et al. 2018; Gao et al. 2018; Kroj et al. 2016; Sarris et al. 2016; Shao et al. 2016; Stein et al. 2018). The most striking findings were the repetitive modular arrangement of NLRs and the discovery of head-to-head paired NLR genes, where one member included an ID (Bailey et al. 2018; Gao et al. 2018; Kroj et al. 2016; Maqbool et al. 2015; Sarris et al. 2016; Shao et al. 2016). The potential use of those IDs as modular building blocks has opened up new possibilities for the engineering of novel resistances to pathogens (Kim et al. 2016; Kourelis et al. 2016; Nishimura et al. 2015). The existing list of IDs, however, likely represents only a glimpse of the true diversity across plants.

The definition of pan-NLR'omes or repertoire of NLR genes across different species, or higher taxonomic groups, has provided estimates of the variation in size of the NLR family (Guo et al. 2011; Stam et al. 2016; Zhang et al. 2010), presence/absence relations (Stam et al. 2016; Zhang et al. 2010), categorical distribution into structural classes across the phylogeny, and diversity of IDs (Kroj et al. 2016; Sarris et al. 2016). Publicly available plant genome annotations have been the foundation of most NLR'ome studies and, their systematic integration has allowed ancestry reconstruction of key NLR lineages and illuminated ancient and recent expansion-contraction events (Shao et al. 2016). In contrast, knowledge of the true diversity of within species pan-NLR'omes is scarce and has so far been derived from only a limited number of individuals, and thus covers a narrow diversity within the population (Guo et al. 2011; Stam et al. 2016; Zhang et al. 2010). Across individuals of the same species, which often has only a single reference genome annotation, the remarkable differences in NLR family size between rice, tomato, and *Arabidopsis thaliana* might be due to low coverage of available genomes, or the impracticable assembly of tandem paralogous genes often found in NLR clusters when short-read sequencing is used under conditions of insufficient depth (Stam et al. 2016; Zhang et al. 2010).

Despite these potential shortcomings, early intraspecific pan-NLR'ome studies revealed patterns of allelic and structural variation consistent with adaptive evolution and balancing selection for subsets of NLR encoding genes (Bakker et al. 2008), fitting a model of co-evolution of host and pathogens. Allelic variation seems to be reflected in many different haplotypes that are found across NLR loci (Cao et al. 2011; Duan et al. 2017). These can include recombination "hotspots" generating NLR clusters (Choi et al. 2016; Guo et al. 2011), and true allelic series (Dodds et al. 2006; Rose et al. 2004). The patterns of presence/absence polymorphisms as well as copy number variation at loci with multiple NLR genes imply that reference genomes may not include representatives of all distinct NLR clades within a species (Cao et al. 2011; Golicz et al. 2016; Guo et al. 2011; Kawakatsu et al. 2016; Li et al. 2014). A major advance in identifying 'missing' NLR genes in a species is resistance gene enrichment sequencing (RenSeq), especially when hybridization based capture of genomic fragments with sequence similarity to known NLR-coding genes is combined with Single-Molecule Real Time sequencing (SMRT RenSeq) (Witek et al. 2016). Our objective was to define the full NLR repertoire and its variability in the reference species A. thaliana, by analyzing a panel of 65 diverse accessions using SMRT RenSeq. We show that we approach saturation of the pan-NLR'ome with this well-chosen set of accessions; we define the core NLR complement of the species and detail novel domain architectures; and we describe presence-absence polymorphisms in non-core NLRs. Together, our work provides a foundation for the identification and cloning of disease resistance genes in more complex species of agronomic importance.

3.4. Results

3.4.1. The Samples

A set of 65 *Arabidopsis thaliana* accessions was selected to explore the diversity of the pan-NLR'ome (fig. 3.1a, table 3.B.2). The selection included 46 accessions from the 1001 Genomes Project, of which 21 belonged to previously identified relict populations characterized by an unusually high amount of genetic diversity (1001_Genomes_Consortium 2016). Additionally, the 19 founder accessions of the MAGIC lines, a resource to dissect the genetics of complex traits were included (Kover et al. 2009; Scarcelli et al. 2007).

3.4.2. NLR Complements

A combination of NLR gene sequence capture (RenSeq) and single-molecule real-time sequencing (SMRT) was used to reconstruct full NLR complements (see section 3.7 for details of bait design, sequencing, assembly, annotation and quality control approaches). In total, we identified 13 167 NLRs, with a range of 167 to 251 genes per accession (fig. 3.1b). Individual accessions had between 47 % and 71 % physically clustered NLR genes (more than one NLR in 200kb of genomic sequence; adapted from Holub (2001)). A particularly interesting class of NLR genes are those in head-to-head orientation (Narusaka et al. 2009; Saucet et al. 2015), and we found 10 to 34 NLRs per accession in such an orientation, or with high sequence similarity to known functional pairs (section 3.7). NLRs were grouped into four classes (TNL, NL, CNL, and RNL) based on canonical protein domains (TIR, NB, CC, RPW8 and LRR). Across all accessions TNLs formed the largest and most size-variable class, followed by NLs, CNLs, and RNLs (fig. 3.1c, fig. 3.A.1). Of the 13 167 NLR genes, 663 contained at least one additional integrated protein domains (ID), in which we found 36 distinct Pfam domains (fig. 3.2b). Individual accessions had 5 to 17 IDs distributed accross 4 to 16 NLR genes, in line with reports for specific accessions (Guo et al. 2011; Shao et al. 2016). This result reveals an unprecedented incidence of previously unreported *A. thaliana* IDs.

3.4.3. NLR Domain Architecture Diversity

We investigated the repertoire of the 36 IDs, since these might function as pathogen effector binding platforms (Cesari et al. 2014; Le Roux et al. 2015; Nishimura et al. 2015; Sarris et al. 2015). 29 of the 36 IDs were already known from other Brassicaceae including *A. thaliana* Col-0 (fig. 3.2a,b; table 3.B.7). Nine of the 36 IDs were reported concordantly in the two major NLR-ID censuses, namely WRKY, PP2, Pkinase, PAH, DUF640, B3, Pkinase_Tyr, PPR_2 and Alliinase_C (Kroj et al. 2016; Sarris et al. 2016). Five of those nine occur in genetically linked paired NLRs in the pan-NLR'ome (pair ratio > 0.5 in fig. 3.2b, see section 3.7). Rediscovery of these nine IDs is of particular relevance, since these are enriched for domains similar to known effector targets (Kroj et al. 2016; Mukhtar et al. 2011; Sarris et al. 2016; Weßling et al. 2014). Our sequencing

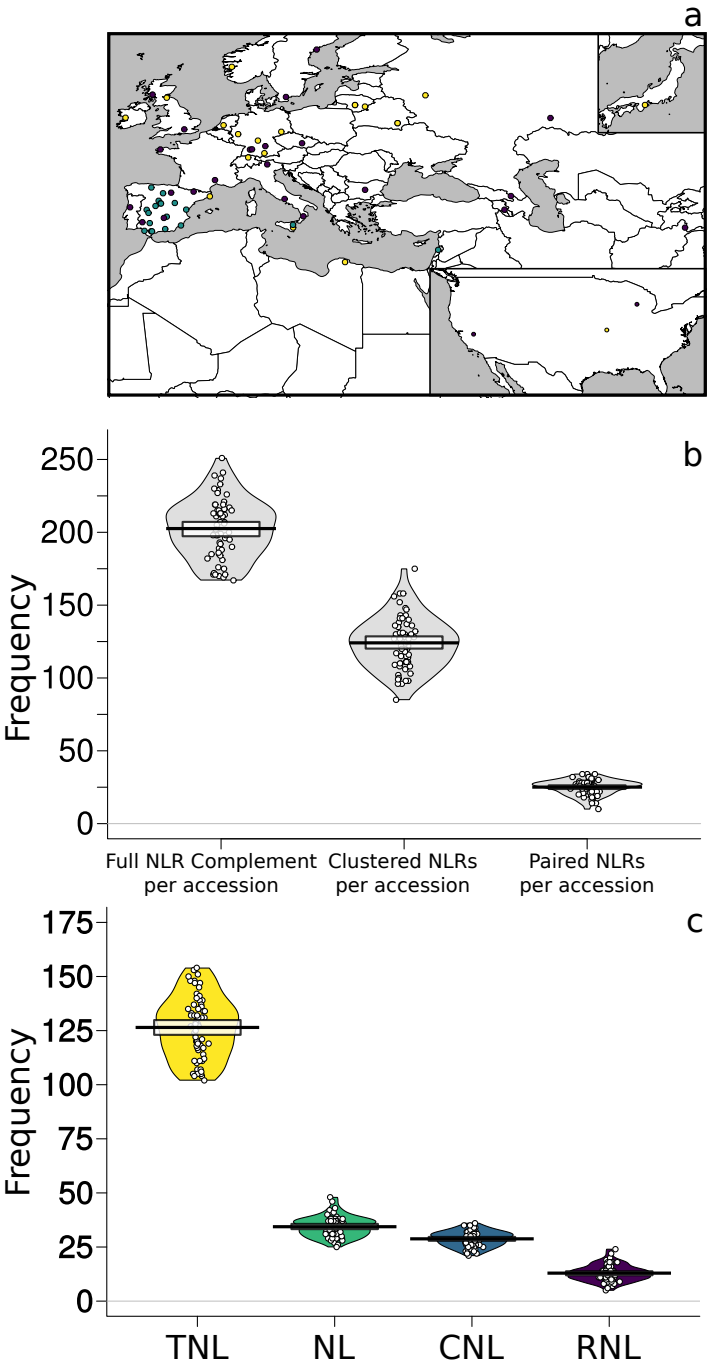


Figure 3.1.: Basic descriptive statistics of the NLR complements
a) World map of *A. thaliana* accessions. 1001G (relics, blue), 1001G (non-relics, purple), MAGIC founders (yellow). b) Accession specific NLR frequencies. Shown are the full NLR complement, clustered NLRs, and paired NLRs. The mean is shown as a solid black horizontal line and the Bayesian 95% Highest density Intervals (HDI: points in the interval have a higher probability than points outside, analogous to 95% confidence intervals) are shown as solid bands around the sample mean. All raw data points are plotted as open circles and the full densities are shown as a bean plot. c) NLR frequency for different structural classes. Mean, HDI, and raw data points are plotted as described.

and annotation effort expands the *A. thaliana* ID repertoire beyond the ten IDs found in the Col-0 reference accession. IDs found in only one gene model did not receive particular attention, as they are conceivably an artefact of our annotation pipeline.

A hallmark of NLR'ome variation across species is the variation in the relative fraction of different domain architectures (Li et al. 2015; Shao et al. 2016). Examining the arrangement of NLR domains in the *A. thaliana* pan-NLR'ome we identified 97 distinct architectures (fig. 3.A.3a,b). Whilst 27 canonical architectures (without IDs) account for the vast majority of the identified NLRs (95% of the pan-NLR'ome; 12 496 NLRs), the remaining 5% (664 NLRs) contain at least one of 36 different IDs (fig. 3.2c). The 97 architectures greatly expand upon the 22 architectures found in the reference Col-0 genome (fig. 3.2d), with most of the new *A. thaliana* architectures containing at least one ID (fig. 3.A.3c). Half of the new *A. thaliana* architectures contain more than one gene (38/75) (fig. 3.2e), of which, 17 are predominantly composed by paired NLRs (pair ratio > 0.5, see section 3.7) and contain at least one ID (fig. 3.2e). About half of the architectures have not been previously described in the Brassicaceae (including *A. thaliana* Col-0) (48/97) (fig. 3.2d). These novel architectures account only for 1.3% of the pan-NLR'ome (175 NLRs), with all but one containing an ID (fig. 3.2d,e). Finally, 12 IDs are repeatedly recruited into different novel architectures (labeled 'novel > known' in fig. 3.2f), reflecting the recycling of a limited set of IDs into new domain arrangements. It is likely that these IDs are derived from proteins repeatedly targeted by pathogen virulence effectors.

3.4.4. The pan-NLR'ome

To begin to understand the diversity of both NLR content and alleles, we grouped sets of homology-related NLRs from different ecotypes. The resulting clusters were termed orthogroups. We clustered 11 497 NLRs into 464 high confidence orthogroups (OGs) (fig. 3.3a), plus 1663 singletons. Ninety-five percent of the OGs could be discovered with 38 randomly chosen accessions (fig. 3.3b). Additional sampling only recovered OGs with three or fewer members, indicating that the pan-NLR'ome we describe is largely, if not completely, saturated. OGs were classified according to size, domain architecture and structural features. We define the core NLR'ome as the 106 OGs found in at least 52 accessions (6,080 genes), 143 OGs found in at least 13, but fewer than 52 accessions as shell (3,932 genes), and the 215 OGs found in 12 or fewer accessions as cloud (1,485 genes) (fig. 3.3a). The majority of OGs, 58%, were TNLs, in concordance with TNLs being the prevalent NLR class in the Brassicaceae (Peele et al. 2014), 22% were CNLs, 7% RNLs, and 13% NLs (fig. 3.3c). TNLs showed a strong tendency towards larger shell and core OGs compared to CNLs (fig. 3.A.5a). Sixtyfour OGs included genetically paired NLRs (see section 3.7), and 28 contained members with an ID, with almost none being present in the cloud NLR'ome (fig. 3.3d). Shell and core OGs contained the vast majority of paired NLRs (98% in 55 OGs, fig. 3.A.6). This shows that conserved NLR pairs are widely distributed in the population and that incorporation of IDs into NLRs is widespread in *A. thaliana*.

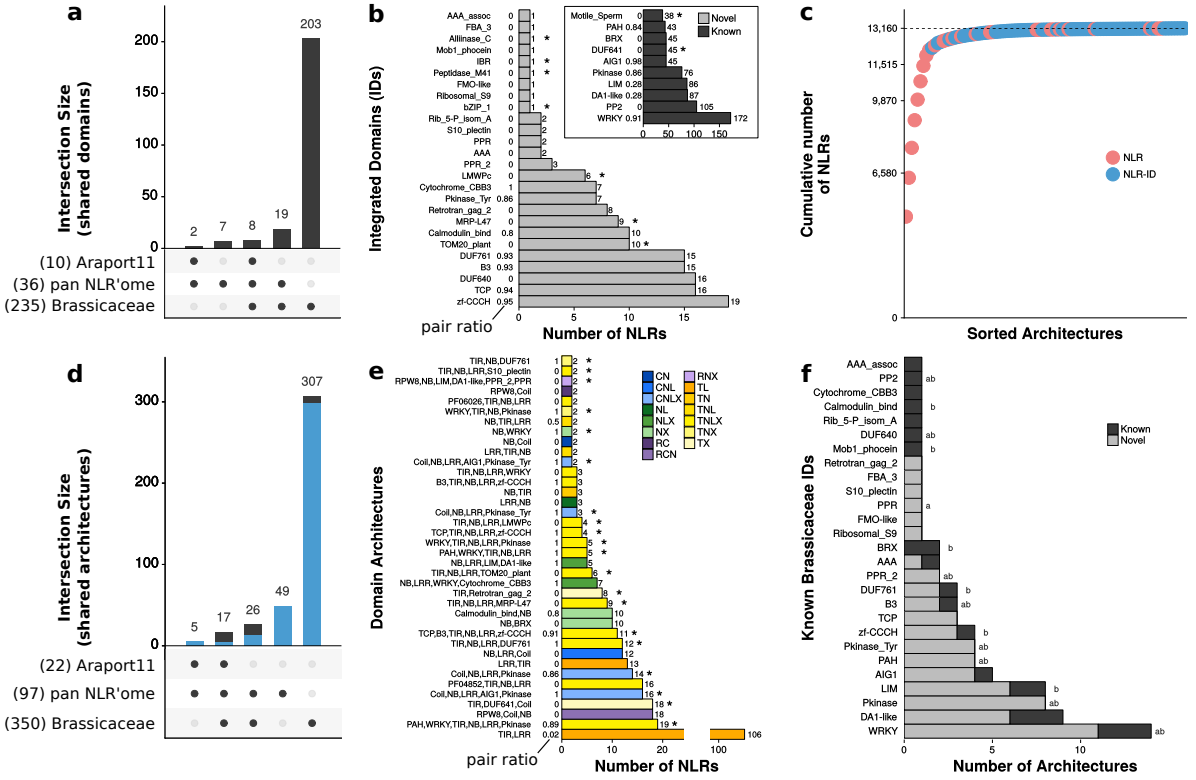


Figure 3.2.: Diversity of IDs and domain architectures in the pan-NLR'ome
a) UpSet intersection of IDs in the pan-NLR'ome with those found in the *A. thaliana* Col-0 reference accession and 19 Brassicaceae genomes. The number of IDs in each set is indicated between parenthesis at the lower left. Set intersections are depicted in the combination matrix in the bottom. Size of intersecting sets is indicated on the vertical bars. b) NLR-ID prevalence. Novel *A. thaliana* IDs are shown in light grey and known IDs in the Araport11 NLR'ome are in black. Pfam30 domain names in y-axis and number of NLRs containing each ID are shown in x-axis. Asterisks show IDs not detected in Brassicaceae NLRs. Numbers next to y-axis show the ratio of paired NLRs among the NLRs containing the ID. c) Cumulative sizes of each of the 97 domain architectures. 27 canonical architectures contain 12496 NLRs (red) and 70 architectures with IDs contain 664 NLRs (blue). d) UpSet intersection plot showing the number of shared architectures between the pan-, Araport11- and Brassicaceae-NLR'ome sets. The number of architectures in each set is shown between parenthesis at the lower left. The number of shared architectures between sets are indicated the respective vertical bars. The blue stacks of the bars indicate the number of ID-containing architectures. e) 38 previously unreported architectures in Col-0 containing more than one gene. 37 architectures represented by single genes are not shown. Domain architectures are shown in the y-axis. The number of NLRs in each architecture is shown in x-axis. Asterisks indicate 20 of the 49 architectures not yet detected in the Brassicaceae family and in the reference Col-0 accession. Numbers next to y-axis show the ratio of paired NLRs divided by the total number of NLRs in each architecture. f) Number of known and novel architectures containing the 27 overlapping Brassicaceae IDs (see panel a). a and b letters at the right of the bars indicate putative integrated decoys as reported by Kroj et al. (2016) and Sarris et al. (2016).

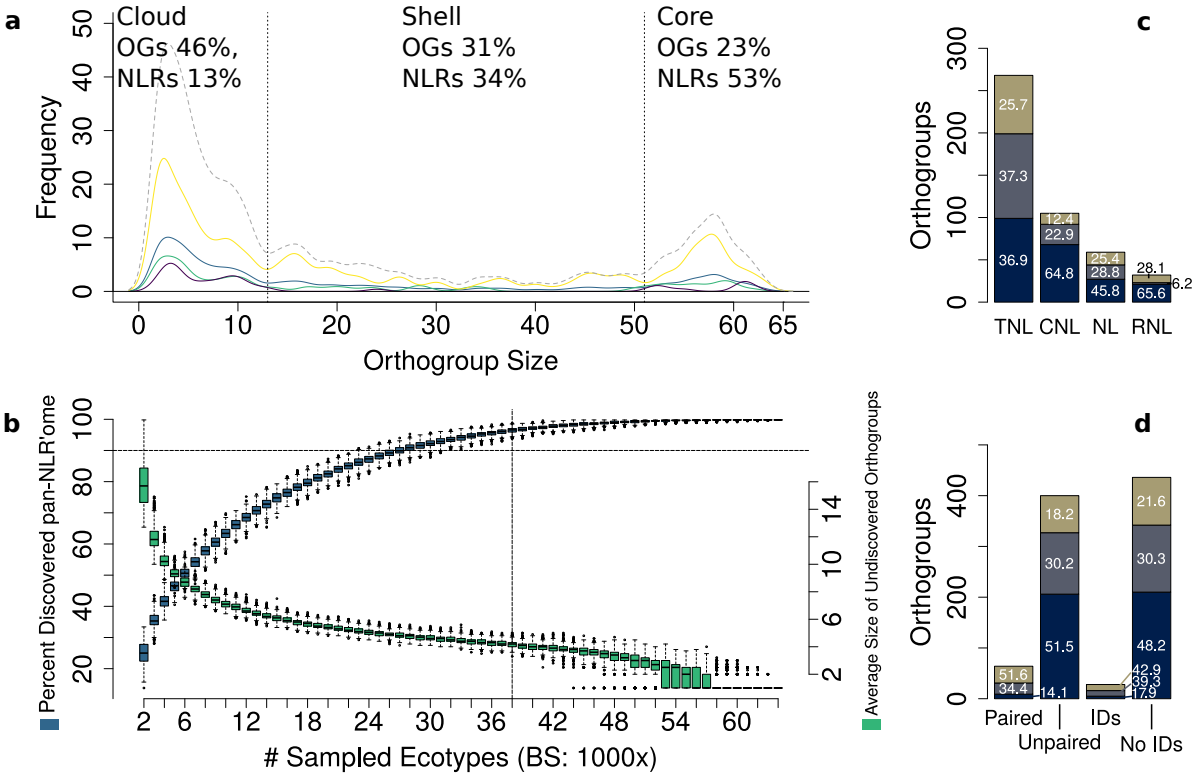


Figure 3.3.: OG sizes, Saturation, Distribution of NLR classes and pairs
a) Orthogroup size distribution. Data is shown separately for the different NLR classes (yellow (TNL), green (NL), blue (CNL), purple (RNL)), and for all NLRs (grey dashed line). The vertical lines at x=13 and x=51 differentiate cloud, shell, and core. b) Saturation of the pan-NLR'ome. The blue boxes show the percentage of the pan-NLR'ome that can be recovered when randomly drawing a fixed number of accessions (1000x bootstrapping). The horizontal dashed line is drawn where 90 % of the pan-NLR'ome is found. The green boxes show for each subset of drawn accessions, the average size of undiscovered orthogroups. The vertical dashed line shows that 95 % of the pan-NLR'ome can be recovered using 38 accessions. c) OG-type specific distribution of NLR classes. Shown is the total number of orthogroups in the Cloud (dark blue), the Shell (grey), and the Core (olive green), and the percentage (text in the bars). d) OG-type specific distribution of paired and unpaired NLRs, and NLRs with and without IDs. Shown is the total number of Orthogroups in the Cloud (dark blue), the Shell (grey), and the Core (olive green), and the percentage (text in the bars).

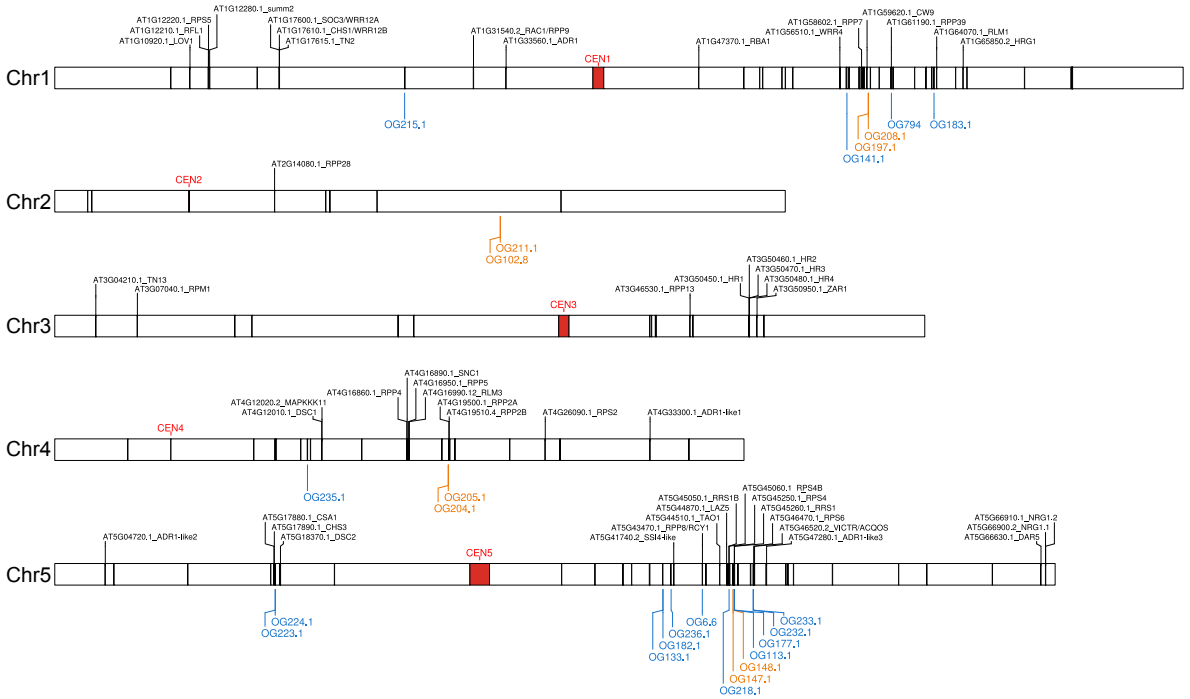


Figure 3.4.: Genetic location of NLRs

Genetic location of NLR-coding genes in the reference *A. thaliana* Col-0 assembly (TAIR9). The 5 *A. thaliana* Col-0 chromosomes are shown in horizontal bars. Centromeres are shown as red regions inside each chromosome. Col-0 NLRs are shown as black line segments inside each chromosome. For simplicity, text labels are shown on top of each chromosome only for functionally defined Col-0 NLRs. Anchored NLR OGs with a threshold of 10 accessions are shown below each chromosome in blue and orange. Novel paired NLR OGs (not previously reported in Col-0) are colored orange, while remaining anchored NLR OGs are shown in blue.

3.4.5. Placement of non-reference OGs

We discovered 296 high confidence OGs without a reference Col-0 allele, with six belonging to the core, 205 to the cloud, and 85 to the shell NLR'ome. In order to anchor these OGs to the reference genome, we asked how often orthogroups co-occurred, using OGs with known location (NLR and non-NLR OGs with a Col-0 reference allele) to anchor contigs with OGs lacking a reference allele. With a minimum threshold of 10 accessions, we derived 42 co-occurrence subnetworks (fig. 3.A.7), anchoring 24 out of 132 OGs present in at least 10 accessions, but missing from the Col-0 reference. Most were anchored to other NLRs (fig. 3.A.7). Newly anchored OGs include one CNL pair and three TNL pairs (fig. 3.4, fig. 3.A.7, and fig. 3.A.9), with one ID-containing sensor-type OG (205.1) arranged in head-to-head orientation to the executor-type OG 204.1 (fig. 3.A.9). The use of annotated non-NLR genes in the assembled contigs allowed us to properly place these novel OGs.

3.4.6. Pan-NLR'ome Diversity

Complementary to defining the diversity of NLR architectures, we assessed sequence diversity and evolutionary forces shaping the pan-NLR'ome. The average nucleotide diversity was similar for CNLs, NLs and TNLs (fig. 3.5a). The same trend was true for haplotype diversity (fig. 3.5c). Nucleotide diversity was lowest in core-type and higher in shell- and cloud-type OGs across TNLs, CNLs and NLs (fig. 3.5a), suggesting that selection is relaxed in OGs with larger presence/absence variation. Interestingly, TNLs and NLs contained a number of shell-type OGs with ultra-low haplotype diversity, suggesting a conserved but rarely encountered selective pressure (there is no correlation between geographic location and the accessions carrying these orthogroups). Both nucleotide and haplotype diversity showed signs of saturation (fig. 3.A.12). The average nucleotide diversity saturated with 32 randomly selected accessions while the haplotype diversity saturated later with 49 accessions. This suggests a prevalence of low frequency haplotypes. Contrasting nucleotide and haplotype diversity in paired, clustered and non-clustered OGs revealed a significant increase in nucleotide diversity in clustered versus non-clustered OGs (fig. 3.A.5c). This prevalence of diversity in clustered NLRs supports the theory of relaxed selection in cases of gene duplication (NLR clusters typically contain arrays of duplicates) (Ohno 1970). Examination of nucleotide diversity across OGs that were each separated into functional protein domains revealed the most diversity in LRRs and steadily decreasing diversity in the other protein domains (LRR > NB > TIR/CC) across all major classes and subclasses (fig. 3.5b).

Our assessment of balancing selection and purifying selection via Tajima's D (TD) showed a similar distribution across TNLs, CNLs and NLs (fig. 3.5d). All classes contain extremes in both directions. Notably, TNLs exhibited a shift towards a lower TD, an effect largely driven by core- and shell-type OGs.

3.4.7. Linking Diversity to Function

Because NLRs that had been experimentally implicated in resistance to biotrophic pathogens showed enhanced diversity, we sorted OGs by resistance to adapted biotrophs (*Hyaloperonospora arabidopsidis*), non-adapted biotrophs (*Albugo candida*), hemibiotrophs (mostly *Pseudomonas* spp.). OGs that provide resistance against adapted biotrophs are significantly more diverse than other categories (fig. 3.6a; ANOVA and Tukey's HSD $p < 0.01$), suggests that host-adapted biotrophic pathogens are driving diversification of NLRs more than other pathogens. That RNL helper NLRs have low diversity is consistent with their requirement to function with several sensor NLRs (Bonardi et al. 2011; Wu et al. 2017).

Among the OGs with the lowest Tajima's D values, a prominent example was RPM1, which confers resistance to hemibiotrophic bacterial pathogens, and for which an ancient, stably balanced presence/absence polymorphism across *A. thaliana* is well established (Stahl et al. 1999). OGs that provide resistance to adapted biotrophs tend to have higher Tajima's D values, indicating that they experience not only diversifying, but also balancing selection. Two OGs with high Tajima's D values are the paired NLRs

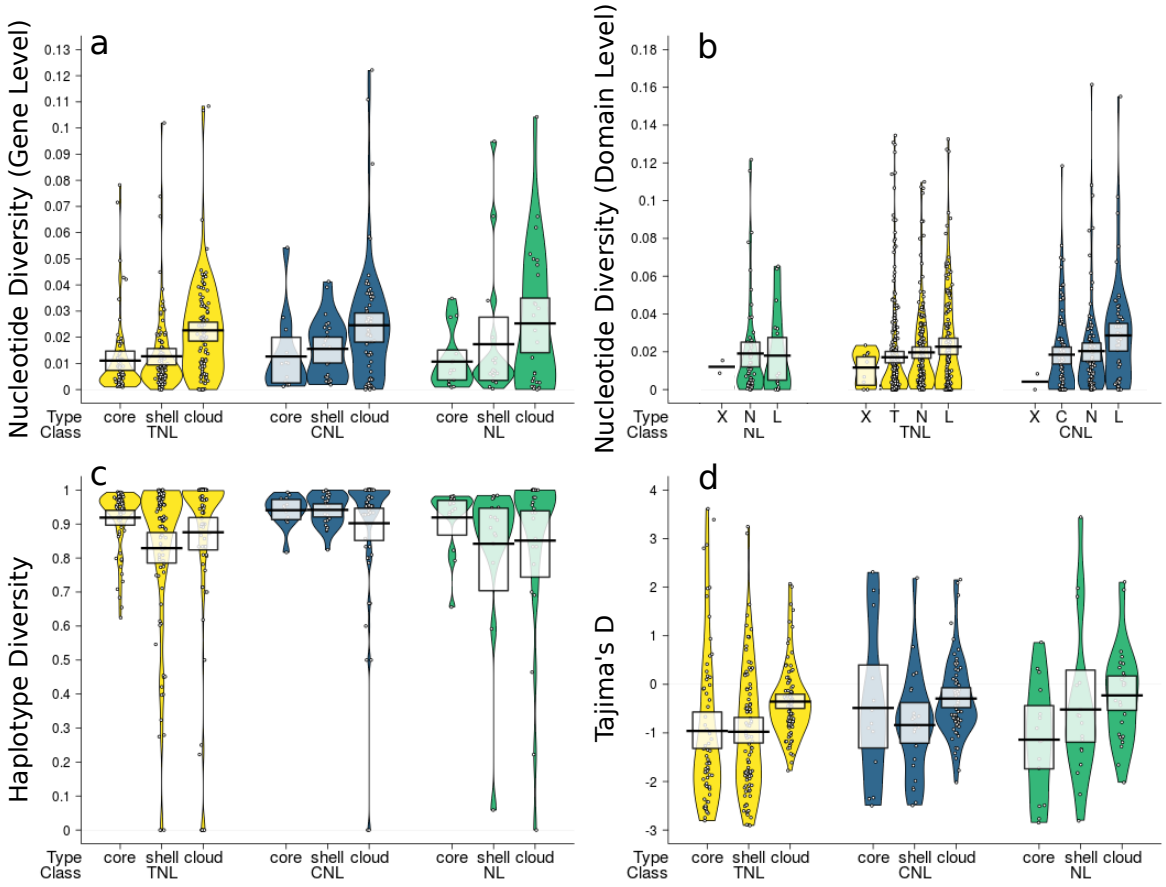


Figure 3.5.: Nucleotide- and haplotype diversity
TNL OGs (yellow), CNL OGs (blue), and NL OGs (green) are shown in core, shell, and cloud. a) Nucleotide diversity on gene level b)Nucleotide diversity on domain level c) Haplotype Diversity d) Tajima's D

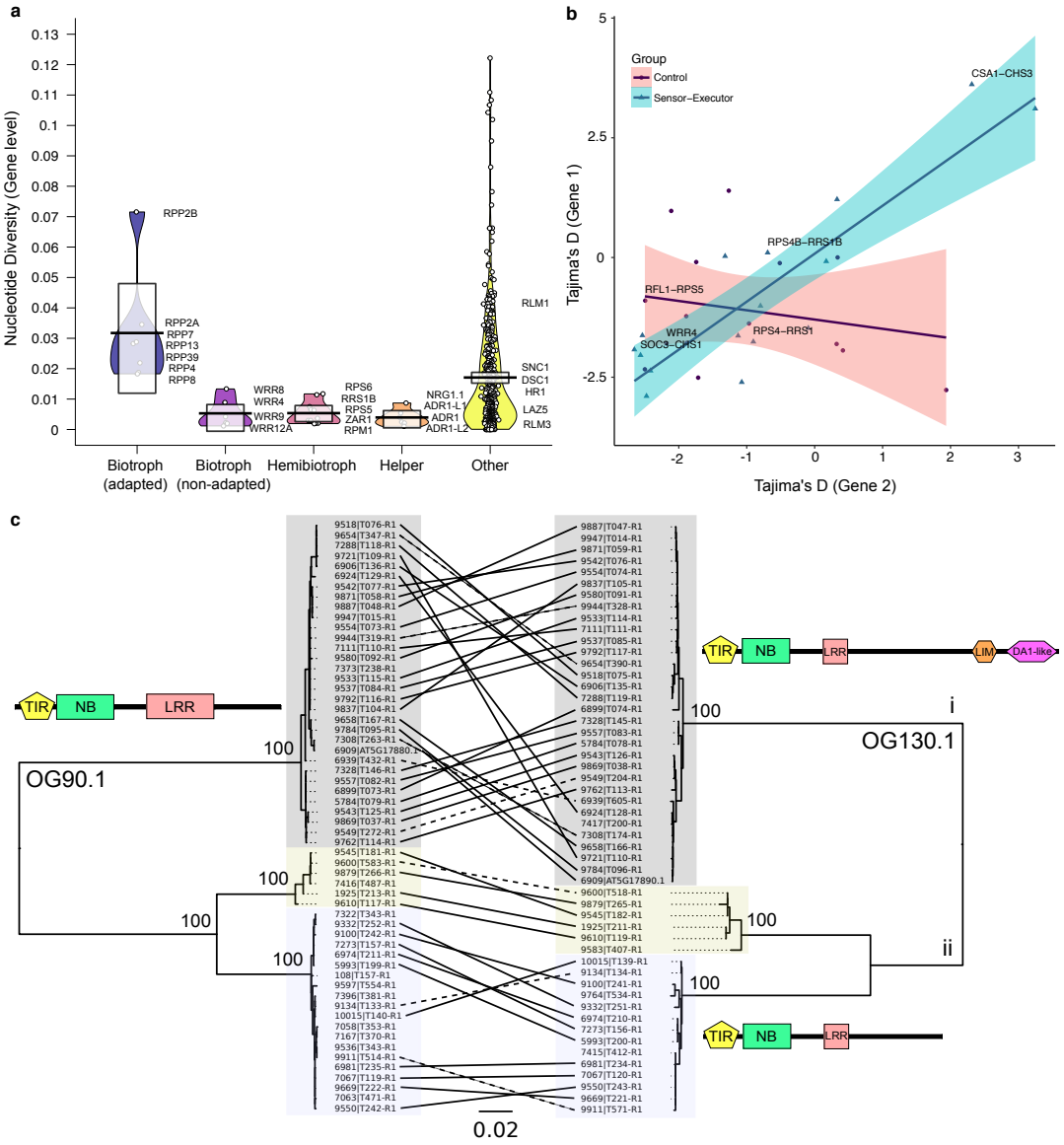


Figure 3.6.: R genes against biotrophic pathogens have enhanced diversity and sensor/executor-like pairs suggest intra-pair co-evolution
a) Nucleotide diversity distributions by functional class. Function was assigned to OGs where possible from five categories: Biotroph (adapted), Biotroph (non-adapted), Hemibiotroph, Helper and Other. b) Plot of sensor/executor and control pairs, where pair member's OG Tajima's D values are plotted against each other, across pairs. Also shown are several exemplary pairs. c) Diagram including the phylogenetic tree of two orthogroups which form a sensor/executor pair (OGs 90.1 and 130.1). Lines drawn between the phylogenies denote where the OG members come from the same accession. Scale schematic diagrams display the architecture of a selected protein from within that clade and are representative of their clades, which include minor variations.

CSA1 (OG91) and CHS3 (OG130). CHS3 featured two very different groups of alleles distinguished by the presence of LIM and DA1-like IDs (Xu et al. 2015). This pattern was perfectly mirrored by the one for CSA1, the paired 'executor' partner NLR of CHS3 (fig. 3.6c). Tajima's D values within sensor-executor pairs encoded in head-to-head orientation were correlated whereas other closely linked NLR genes or random pairs were not (fig. 3.6b, table 3.B.5).

3.5. Discussion

We defined the full species repertoire of the gene family that encodes NLR immune receptors in the model plant *A. thaliana*. Our most important finding is perhaps that the pan-NLR'ome inventory became already saturated with ∼40 accessions randomly selected from the 65 accessions we analyzed. Before our work, it was known that there was excessive variation at some NLR loci, such that in the small number of accessions in which the relevant genomic region was analyzed in detail, every accession was very different (Noel 1999; Rose et al. 2004), suggesting that there were dozens, if not hundreds of substantially different alleles. The fact that our pan-NLR'ome saturates with ∼40 accessions indicates that the number of divergent alleles is not unlimited. It also provides some guidance for future efforts in other species. Among functionally annotated genes, we found the highest sequence diversity in NLR resistance genes whose products recognize evolutionarily adapted biotrophic pathogens.

We have also found an astonishing diversity of IDs, which allow hosts to rapidly accrue the ability to recognize the biochemical action of pathogen effector proteins. ID containing NLRs that have been functionally characterized are all found in paired orientation. In these pairs, the ID member functions as pathogen sensor, and the other member as signaling executor (Cesari et al. 2014; Le Roux et al. 2015; Narusaka et al. 2009; Sarris et al. 2015; Xu et al. 2015; Zhang et al. 2017), with both members contributing to repression and activation of NLR signaling (Ma et al. 2018). The correlation between Tajima's D values of such paired NLRs support a co-evolutionary scenario whereby mutations into the sensor component lead to compensatory changes in the executor, or vice versa.

However, half of the 22 most commonly found IDs did not occur in an arrangement indicative of sensor/executor pairs. An open question is whether these function with unlinked executor partners, or whether they can function as dual sensor/executor proteins. Within the *A. thaliana* pan-NLR'ome, we identified three key families of defense-related TCP, WRKY and CBP60 transcription factors, represented as IDs in sensors of the class defined by RRS1. TCP domains are particularly interesting, as TCP transcription factors are preferentially targeted by pathogen effectors from divergently evolved pathogens (Mukhtar et al. 2011; Sugio et al. 2014; Weßling et al. 2014; Yang et al. 2017). The TCP domain may open a new avenue to engineering of NLR specificity, through TCP swap or inclusion of known effector-interacting platforms from TCP14 (Yang et al. 2017), as recently proven with protease cleavage site swaps (Helm et al. 2018; Kim et al. 2016).

3.6. Bibliography

References

1001_Genomes_Consortium (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". In: *Cell* 166.2. URL: http://dx.doi.org/10.1016/j.cell.2016.05.063.

Bailey, Paul C., Christian Schudoma, William Jackson, Erin Baggs, Gulay Dagdas, Wilfried Haerty, Matthew Moscou, and Ksenia V. Krasileva (2018). "Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions". In: *Genome Biology* 19.1.

Bakker, Erica G., M. Brian Traw, Christopher Toomajian, Martin Kreitman, and Joy Bergelson (2008). "Low levels of polymorphism in genes that control the activation of defense response in Arabidopsis thaliana". In: *Genetics* 178.4.

Bernoux, Maud, Thomas Ve, Simon Williams, Christopher Warren, Danny Hatters, Eugene Valkov, Xiaoxiao Zhang, Jeffrey G. Ellis, Bostjan Kobe, and Peter N. Dodds (2011). "Structural and functional analysis of a plant resistance protein TIR domain reveals interfaces for self-association, signaling, and autoregulation". In: *Cell Host and Microbe* 9.3.

Bonardi, V., S. Tang, A. Stallmann, M. Roberts, K. Cherkis, and Jeffery L. Dangl (2011). "Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors". In: *Proceedings of the National Academy of Sciences* 108.39. arXiv: `arXiv:1408.1149`. URL: `http://www.pnas.org/cgi/doi/10.1073/pnas.1113726108`.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, Xi Wang, Felix Ott, Jonas Müller, Carlos Alonso-Blanco, Karsten Borgwardt, Karl J Schmid, and Detlef Weigel (Oct. 2011). "Whole-genome sequencing of multiple Arabidopsis thaliana populations." In: *Nature genetics* 43.10. URL: `http://dx.doi.org/10.1038/ng.911`.

Catanzariti, Ann-Maree, Peter N Dodds, Thomas Ve, Bostjan Kobe, Jeffrey G Ellis, and Brian J Staskawicz (2010). "The AvrM Effector from Flax Rust Has a Structured C-Terminal Domain and Interacts Directly with the M Resistance Protein". In: *Mol. Plant. Microbe. Interact.* 23.1.

Cesari, Stella, Maud Bernoux, Philippe Moncuquet, Thomas Kroj, and Peter N. Dodds (2014). "A novel conserved mechanism for plant NLR protein pairs: the "integrated decoy" hypothesis". In: *Frontiers in Plant Science* 5.November. URL: `http://journal.frontiersin.org/article/10.3389/fpls.2014.00606/abstract`.

Choi, Kyuha, Carsten Reinhard, Heïdi Serra, Piotr A Ziolkowski, Charles J. Underwood, Xiaohui Zhao, Thomas J. Hardcastle, Nataliya E. Yelina, Catherine Griffin, Matthew Jackson, Christine Mézard, Gil McVean, Gregory P. Copenhaver, and Ian R Henderson (2016). "Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes". In: *PLoS Genetics* 12.7.

Dangl, Jeffery L., Diana M. Horvath, and Brian J. Staskawicz (2013). *Pivoting the plant immune system from dissection to deployment.* arXiv: `NIHMS150003`.

Dodds, Peter N, Gregory J Lawrence, A.-M. Catanzariti, Trazel Teh, C.-I. A Wang, Michael A Ayliffe, Bostjan Kobe, and Jeffrey G Ellis (2006). "Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes". In: *Proceedings of the National Academy of Sciences* 103.23. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0602577103.

Duan, Naibin, Yang Bai, Honghe Sun, Nan Wang, Yumin Ma, Mingjun Li, Xin Wang, Chen Jiao, Noah Legall, Linyong Mao, Sibao Wan, Kun Wang, Tianming He, Shouqian Feng, Zongying Zhang, Zhiquan Mao, Xiang Shen, Xiaoliu Chen, Yuanmao Jiang, Shujing Wu, Chengmiao Yin, Shunfeng Ge, Long Yang, Shenghui Jiang, Haifeng Xu, Jingxuan Liu, Deyun Wang, Changzhi Qu, Yicheng Wang, Weifang Zuo, Li Xiang, Chang Liu, Daoyuan Zhang, Yuan Gao, Yimin Xu, Kenong Xu, Thomas Chao, Gennaro Fazio, Huairui Shu, Gan Yuan Zhong, Lailiang Cheng, Zhangjun Fei, and Xuesen Chen (2017). "Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement". In: *Nature Communications* 8.1. URL: http://dx.doi.org/10.1038/s41467-017-00336-7.

Gao, Yuxia, Wenqiang Wang, Tian Zhang, Zhen Gong, Huayao Zhao, and Guan-Zhu Han (2018). "Out of Water: The Origin and Early Diversification of Plant R-Genes". In: *Plant Physiology* 31701091. URL: http://www.plantphysiol.org/lookup/doi/10.1104/pp.18.00185.

Golicz, Agnieszka A., Philipp E. Bayer, Guy C. Barker, Patrick P. Edger, HyeRan Kim, Paula A. Martinez, Chon Kit Kenneth Chan, Anita Severn-Ellis, W. Richard McCombie, Isobel A. P. Parkin, Andrew H. Paterson, J. Chris Pires, Andrew G. Sharpe, Haibao Tang, Graham R. Teakle, Christopher D. Town, Jacqueline Batley, David Edwards, S. Liu, I. Parkin, M. Morgante, X. Gan, J. Cao, A. A. Golicz, J. Batley, D. Edwards, W. Yao, C. N. Hirsch, Y.-H. Li, H. Tettelin, S. J. Bush, M. Schatz, R. E. Mills, B. Weckselblatt, M. K. Rudd, J. Zhang, T. Zuo, T. Peterson, K. Song, T. C. Osborn, P. H. Williams, X. Xu, L. K. McHale, M. A. Lysak, M. A. Lysak, M. A. Koch, A. Pecinka, I. Schubert, B. Chalhoub, B. C. Meyers, A. Kozik, A. Griego, H. Kuang, R. W. Michelmore, K. Lin, T. C. Osborn, M. Tadege, M. E. Schranz, K. Okazaki, J. Zhao, S.-Y Kim, D. Xiao, S. Ridge, P. H. Brown, V. Hecht, R. G. Driessen, J. L. Weller, M. M. Kushad, D. J. Kliebenstein, V. M. Lambrix, M. Reichelt, J. Gershenzon, T. Mitchell-Olds, J. A. Hofberger, E. Lyons, P. P. Edger, J. C. Pires, M. E. Schranz, P. P. Edger, J. Zhang, G. Li, C. F. Quiros, N. M. Springer, R. A. Swanson-Wagner, K. Schneeberger, B. Langmead, S. L. Salzberg, A. V. Zimin, A. M. Bolger, M. Lohse, B. Usadel, C. Camacho, T. Arias, M. A. Beilstein, M. Tang, M. R. McKain, J. C. Pires, C. Holt, M. Yandell, I. Korf, M. Stanke, S. R. Eddy, J. Piriyapongsa, M. T. Rutledge, S. Patel, M. Borodovsky, I. K. Jordan, A. V. McDonnell, T. Jiang, A. E. Keating, B. Berger, E. B. Holub, E. Richly, J. Kurth, D. Leister, K. Howe, A. Bateman, R. Durbin, H. Li, A. Golicz, L. Li, C. J. Stoeckert, D. S. Roos, H. Tettelin, D. Riley, C. Cattuto, D. Medini, A. Rimmer, V. Obenchain, A. Stamatakis, A. Conesa, A. Alexa, J. Rahnenführer, T. Lengauer, X. Wang, K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, T. Sotelo, P. Soengas, P.

Velasco, V. M. Rodríguez, and M. E. Cartea (Nov. 2016). "The pangenome of an agronomically important crop plant Brassica oleracea". In: *Nature Communications* 7. URL: http://www.nature.com/doifinder/10.1038/ncomms13390.

Guo, Ya-Long, Joffrey Fitz, Korbinian Schneeberger, Stephan Ossowski, Jun Cao, and Detlef Weigel (Oct. 2011). "Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis." In: *Plant physiology* 157.2. URL: http://www.plantphysiol.org/content/157/2/757%20http://www.plantphysiol.org/content/157/2/757/suppl/DC1.

Helm, Matthew, Mingsheng Qi, Shayan Sarkar, Haiyue Yu, Steven A. Whitham, and Roger W. Innes (2018). "Engineering a Decoy Substrate in Soybean to Enable Recognition of the Soybean Mosaic Virus NIa Protease". In: *bioarxiv*.

Holub, Eric B (2001). "The arms race is ancient history in Arabidopsis, the wildflower". In: *Nature* 2.7.

Jones, Jonathan D G and Jeffery L. Dangl (Nov. 2006). "The plant immune system." In: *Nature* 444.7117. URL: http://dx.doi.org/10.1038/nature05286.

Jones, Jonathan D G, Russell E Vance, and Jeffery L Dangl (2016). "Intracellular innate immune surveillance devices in plants and animals". In: *Science* 354.6316.

Kawakatsu, Taiji, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, Joseph R Nery, Cesar Barragan, Yupeng He, Huaming Chen, Manu Dubin, Cheng Ruei Lee, Congmao Wang, Felix Bemm, Claude Becker, Ryan O'Neil, Ronan C O'Malley, Danjuma X Quarless, Carlos Alonso-Blanco, Jorge Andrade, Felix Bemm, Joy Bergelson, Karsten Borgwardt, Eunyoung Chae, Todd Dezwaan, Wei Ding, Joseph R Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G Grimm, Angela Hancock, Stefan R Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Chen Ruei Lee, Dazhe Meng, Todd P Michael, Richard Mott, Ni Wayan W. Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Novikova, F. Xavier Pico, Alexander Platzer, Fernando A Rabanal, Alex Rodriguez, Beth A Rowan, Patrice A. Salome, Karl Schmid, Robert J Schmitz, Ü Seren, Felice Gianluca G. Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, Xuefeng Zhou, Nicholas J. Schork, Detlef Weigel, Magnus Nordborg, and Joseph R. Ecker (2016). "Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions". In: *Cell* 166.2.

Kim, Sang Hee, Dong Qi, Tom Ashfield, Matthew Helm, and Roger W. Innes (2016). "Using decoys to expand the recognition specificity of a plant disease resistance protein". In: *Science* 351.6274. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.aad3436.

Kourelis, Jiorgos, R A L van der Hoorn, and Daniela J Sueldo (2016). "Decoy Engineering: The Next Step in Resistance Breeding". In: *Trends Plant Sci.* 21.5.

Kourelis, Jiorgos and Renier A. L. van der Hoorn (2018). "Defended to the Nines: 25 years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function". In: *The Plant Cell* 30.February. URL: http://www.plantcell.org/lookup/doi/10.1105/tpc.17.00579.

Kover, Paula X., William Valdar, Joseph Trakalo, Nora Scarcelli, Ian M. Ehrenreich, Michael D. Purugganan, Caroline Durrant, and Richard Mott (2009). "A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana". In: *PLoS Genetics* 5.7. arXiv: 15334406.

Krasileva, Ksenia V., D. Dahlbeck, and B. J. Staskawicz (2010). "Activation of an Arabidopsis Resistance Protein Is Specified by the in Planta Association of Its Leucine-Rich Repeat Domain with the Cognate Oomycete Effector". In: *the Plant Cell Online* 22.7. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.110.075358.

Kroj, Thomas, Emilie Chanclud, Corinne Michel-Romiti, Xavier Grand, and Jean Benoit Morel (2016). "Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread". In: *New Phytologist* 210.2.

Le Roux, Clémentine, Gaëlle Huet, Alain Jauneau, Laurent Camborde, Dominique Trémousaygue, Alexandra Kraut, Binbin Zhou, Marie Levaillant, Hiroaki Adachi, Hirofumi Yoshioka, Sylvain Raffaele, Richard Berthomé, Yohann Couté, Jane E. Parker, and Laurent Deslandes (2015). "A Receptor Pair with an Integrated Decoy Converts Pathogen Disabling of Transcription Factors to Immunity". In: *Cell* 161.5. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867415004420.

Li, Xin, Paul Kapos, and Yuelin Zhang (2015). "NLRs in plants". In: *Current Opinion in Immunology* 32. URL: http://dx.doi.org/10.1016/j.coi.2015.01.014.

Li, Ying-hui, Guangyu Zhou, Jianxin Ma, Wenkai Jiang, Long-guo Jin, Zhouhao Zhang, Yong Guo, Jinbo Zhang, Yi Sui, Liangtao Zheng, Shan-shan Zhang, Qiyang Zuo, Xue-hui Shi, Yan-fei Li, Wan-ke Zhang, Yiyao Hu, Guanyi Kong, Hui-long Hong, Bing Tan, Jian Song, Zhang-xiong Liu, Yaoshen Wang, Hang Ruan, Carol K L Yeung, Jian Liu, Hailong Wang, Li-juan Zhang, Rong-xia Guan, Ke-jing Wang, Wen-bin Li, Shou-yi Chen, Ru-zhen Chang, Zhi Jiang, Scott a Jackson, Ruiqiang Li, and Li-juan Qiu (2014). "De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits". In: *Nature Biotechnology* 32.10. URL: http://dx.doi.org/10.1038/nbt.2979.

Ma, Yan, Hailong Guo, Lanxi Hu, Paula Pons Martinez, Panagiotis N Moschou, Volkan Cevik, Pingtao Ding, Zane Duxbury, Panagiotis F Sarris, and Jonathan D G Jones (2018). "Distinct modes of derepression of an Arabidopsis immune receptor complex by two different bacterial effectors". In: *Proc. Natl. Acad. Sci. U. S. A.* 115.41.

Mackey, David, Ben F. Holt, Aaron Wiig, and Jeffery L. Dangl (2002). "RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis". In: *Cell*.

Maekawa, Takaki, Wei Cheng, Laurentiu N. Spiridon, Armin Töller, Ewa Lukasik, Yusuke Saijo, Peiyuan Liu, Qian Hua Shen, Marius A. Micluta, Imre E. Somssich, Frank L W Takken, Andrei Jose Petrescu, Jijie Chai, and Paul Schulze-Lefert (2011). "Coiled-coil domain-dependent homodimerization of intracellular barley immune receptors defines a minimal functional module for triggering cell death". In: *Cell Host and Microbe* 9.3.

Maqbool, A., H. Saitoh, M. Franceschetti, C. E.M. Stevenson, A. Uemura, H. Kanzaki, S. Kamoun, R. Terauchi, and M. J. Banfield (2015). "Structural basis of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor". In: *eLife*.

Mukhtar, M Shahid, Anne-Ruxandra Carvunis, Matija Dreze, Petra Epple, Jens Steinbrenner, Jonathan Moore, Murat Tasan, Mary Galli, Tong Hao, Marc T Nishimura, Samuel J Pevzner, Susan E Donovan, Lila Ghamsari, Balaji Santhanam, Viviana Romero, Matthew M Poulin, Fana Gebreab, Bryan J Gutierrez, Stanley Tam, Dario Monachello, Mike Boxem, Christopher J Harbort, Nathan McDonald, Lantian Gai, Huaming Chen, Yijian He, European Union Effectoromics Consortium, Jean Vandenhaute, Frederick P Roth, David E Hill, Joseph R Ecker, Marc Vidal, Jim Beynon, Pascal Braun, and Jeffery L Dangl (2011). "Independently evolved virulence effectors converge onto hubs in a plant immune system network". In: *Science* 333.6042.

Narusaka, Mari, Ken Shirasu, Yoshiteru Noutoshi, Yasuyuki Kubo, Tomonori Shiraishi, Masaki Iwabuchi, and Yoshihiro Narusaka (2009). "RRS1 and RPS4 provide a dual Resistance-gene system against fungal and bacterial pathogens". In: *Plant Journal* 60.2.

Nishimura, Marc T, Ryan G Anderson, Karen A Cherkis, Terry F Law, Qingli L Liu, and Mischa Machius (2017). "TIR-only protein RBA1 recognizes a pathogen effector to regulate cell death in Arabidopsis". In: *Proceedings of the National Academy of Sciences* 114.10. URL: `http://www.pnas.org/content/114/10/E2053.abstract%20http://www.pnas.org/content/114/10/E2053.full.pdf`.

Nishimura, Marc T, Freddy Monteiro, and Jeffery L Dangl (2015). "Treasure your exceptions: unusual domains in immune receptors reveal host virulence targets". In: *Cell* 161.5.

Noel, L. (1999). "Pronounced Intraspecific Haplotype Divergence at the RPP5 Complex Disease Resistance Locus of Arabidopsis". In: *the Plant Cell Online* 11.11. URL: `http://www.plantcell.org/cgi/doi/10.1105/tpc.11.11.2099`.

Ohno, Susumo (1970). *Evolution by Gene Duplication*.

Peele, Hanneke M, Na Guan, Johan Fogelqvist, and Christina Dixelius (2014). "Loss and retention of resistance genes in five species of the Brassicaceae family". In: *BMC Plant Biology* 14. URL: `http://www.biomedcentral.com/1471-2229/14/298`.

Qi, D., B. J. DeYoung, and R. W. Innes (2012). "Structure-Function Analysis of the Coiled-Coil and Leucine-Rich Repeat Domains of the RPS5 Disease Resistance Pro-

tein". In: *Plant Physiology* 158.4. URL: http://www.plantphysiol.org/cgi/doi/10.1104/pp.112.194035.

Qi, D., U. Dubiella, S. H. Kim, D. I. Sloss, R. H. Dowen, J. E. Dixon, and R. W. Innes (2014). "Recognition of the Protein Kinase AVRPPHB SUSCEPTIBLE1 by the Disease Resistance Protein RESISTANCE TO PSEUDOMONAS SYRINGAE5 Is Dependent on S-Acylation and an Exposed Loop in AVRPPHB SUSCEPTIBLE1". In: *PLANT PHYSIOLOGY*.

Rose, Laura E, Peter D Bittner-eddy, Charles H Langley, Eric B Holub, Richard W Michelmore, and Jim L Beynon (2004). "The Maintenance of Extreme Amino Acid Diversity at the Disease Resistance Gene, RPP13, in Arabidopsis thaliana". In: *Genetics* 166.2.

Sarris, Panagiotis F., Zane Duxbury, Sung Un Huh, Yan Ma, Cécile Segonzac, Jan Sklenar, Paul Derbyshire, Volkan Cevik, Ghanasyam Rallapalli, Simon B. Saucet, Lennart Wirthmueller, Frank L.H. Menke, Kee Hoon Sohn, and Jonathan D.G. Jones (2015). "A plant immune receptor detects pathogen effectors that target WRKY transcription factors". In: *Cell* 161.5.

Sarris, Panagiotis F, Volkan Cevik, Gulay Dagdas, Jonathan D G Jones, and Ksenia V. Krasileva (2016). "Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens." In: *BMC biology* 14.1. URL: http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-016-0228-7.

Saucet, Simon B., Yan Ma, Panagiotis F. Sarris, Oliver J. Furzer, Kee Hoon Sohn, and Jonathan D.G. Jones (2015). "Two linked pairs of Arabidopsis TNL resistance genes independently confer recognition of bacterial effector AvrRps4". In: *Nature Communications* 6. URL: http://dx.doi.org/10.1038/ncomms7338.

Scarcelli, N., J. M. Cheverud, B. A. Schaal, and P. X. Kover (2007). "Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus". In: *Proceedings of the National Academy of Sciences* 104.43. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0708209104.

Shao, Zhu-Qing, Jia-Yu Xue, Ping Wu, Yan-Mei Zhang, Yue Wu, Yue-Yu Hang, Bin Wang, and Jian-Qun Chen (Apr. 2016). "Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns." In: *Plant physiology* 170.4. URL: http://www.ncbi.nlm.nih.gov/pubmed/26839128%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4825152.

Stahl, E A, G Dwyer, R Mauricio, M Kreitman, and J Bergelson (Aug. 1999). "Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis". In: *Nature* 400.6745.

Stam, Remco, Daniela Scheikl, and Aurelien Tellier (2016). "Pooled Enrichment Sequencing Identifies Diversity and Evolutionary Pressures at NLR Resistance Genes within a Wild Tomato Population". In: *Genome biology and evolution* 8.5.

Stein, Joshua C., Yeisoo Yu, Dario Copetti, Derrick J. Zwickl, Li Zhang, Chengjun Zhang, Kapeel Chougule, Dongying Gao, Aiko Iwata, Jose Luis Goicoechea, Sharon Wei, Jun Wang, Yi Liao, Muhua Wang, Julie Jacquemin, Claude Becker, Dave Kudrna, Jianwei Zhang, Carlos E.M. Londono, Xiang Song, Seunghee Lee, Paul Sanchez, Andrea Zuccolo, Jetty S.S. Ammiraju, Jayson Talag, Ann Danowitz, Luis F. Rivera, Andrea R. Gschwend, Christos Noutsos, Cheng Chieh Wu, Shu Min Kao, Jhih Wun Zeng, Fu Jin Wei, Qiang Zhao, Qi Feng, Moaine El Baidouri, Marie Christine Carpentier, Eric Lasserre, Richard Cooke, Daniel Da Rosa Farias, Luciano Carlos Da Maia, Railson S. Dos Santos, Kevin G. Nyberg, Kenneth L. McNally, Ramil Mauleon, Nickolai Alexandrov, Jeremy Schmutz, Dave Flowers, Chuanzhu Fan, Detlef Weigel, Kshirod K. Jena, Thomas Wicker, Mingsheng Chen, Bin Han, Robert Henry, Yue Ie C. Hsing, Nori Kurata, Antonio Costa De Oliveira, Olivier Panaud, Scott A. Jackson, Carlos A. Machado, Michael J. Sanderson, Manyuan Long, Doreen Ware, and Rod A. Wing (2018). "Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza". In: *Nature Genetics* 50.2. URL: http://dx.doi.org/10.1038/s41588-018-0040-0.

Sugio, Akiko, Allyson M MacLean, and Saskia A Hogenhout (2014). "The small phytoplasma virulence effector SAP11 contains distinct domains required for nuclear targeting and CIN-TCP binding and destabilization". In: *New Phytol.* 202.3.

Takken, Frank L W and Aska Goverse (2012). "How to build a pathogen detector: Structural basis of NB-LRR function". In: *Current Opinion in Plant Biology* 15.4. URL: http://dx.doi.org/10.1016/j.pbi.2012.05.001.

Uehling, Jessie, Aurelie Deveau, and Mathieu Paoletti (2017). "Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems". In: *PLOS Pathogens* 13.10.

Van der Biezen, E A and Jonathan D G Jones (Dec. 1998). "Plant disease-resistance proteins and the gene-for-gene concept." In: *Trends in biochemical sciences* 23.12. URL: http://www.ncbi.nlm.nih.gov/pubmed/9868361.

Wang, Guoxun, Brice Roux, Feng Feng, Endrick Guy, Lin Li, Nannan Li, Xiaojuan Zhang, Martine Lautier, Marie Françoise Jardinaud, Matthieu Chabannes, Matthieu Arlat, She Chen, Chaozu He, Laurent D. Noël, and Jian Min Zhou (2015). "The Decoy Substrate of a Pathogen Effector and a Pseudokinase Specify Pathogen-Induced Modified-Self Recognition and Immunity in Plants". In: *Cell Host and Microbe*.

Weßling, Ralf, Petra Epple, Stefan Altmann, Yijian He, Li Yang, Stefan R Henz, Nathan McDonald, Kristin Wiley, Kai Christian Bader, Christine Gläßer, M Shahid Mukhtar, Sabine Haigis, Lila Ghamsari, Amber E Stephens, Joseph R Ecker, Marc Vidal, Jonathan D G Jones, Klaus F X Mayer, Emiel van Themaat, Detlef Weigel, Paul

Schulze-Lefert, Jeffery L Dangl, Ralph Panstruga, and Pascal Braun (Sept. 2014). "Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life". In: *Cell Host Microbe* 16.3.

Witek, Kamil, Florian Jupe, Agnieszka I Witek, David Baker, Matthew D Clark, and Jonathan D G Jones (Apr. 2016). "Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing". In: *Nature Biotechnology* advance on. URL: http://dx.doi.org/10.1038/nbt.3540.

Wu, Chih-Hang, Ahmed Abd-El-Haliem, Tolga O. Bozkurt, Khaoula Belhaj, Ryohei Terauchi, Jack H. Vossen, and Sophien Kamoun (2017). "NLR network mediates immunity to diverse plant pathogens". In: *Proceedings of the National Academy of Sciences* 114.30. arXiv: 1504.00980. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1702041114.

Wu, Chih-Hang, Ksenia V. Krasileva, Mark J. Banfield, Ryohei Terauchi, and Sophien Kamoun (2015). "The "sensor domains" of plant NLR proteins: more than decoys?" In: *Frontiers in Plant Science* 6.March. URL: http://journal.frontiersin.org/Article/10.3389/fpls.2015.00134/abstract.

Xu, Fang, Chipan Zhu, Volkan Cevik, Kaeli Johnson, Yanan Liu, Kee Sohn, Jonathan D Jones, Eric B Holub, and Xin Li (2015). "Autoimmunity conferred by chs3-2D relies on CSA1, its adjacent TNL-encoding neighbour". In: *Sci. Rep.* 5.

Yang, Li, Paulo José Pereira Lima Teixeira, Surojit Biswas, Omri M. Finkel, Yijian He, Isai Salas-Gonzalez, Marie E. English, Petra Epple, Piotr Mieczkowski, and Jeffery L Dangl (2017). "Pseudomonas syringae Type III Effector HopBB1 Promotes Host Transcriptional Repressor Degradation to Regulate Phytohormone Responses and Virulence". In: *Cell Host and Microbe* 21.2.

Zhang, Meiping, Yen Hsuan Wu, Mi Kyung Lee, Yun Hua Liu, Ying Rong, Teofila S. Santos, Chengcang Wu, Fangming Xie, Randall L. Nelson, and Hong Bin Zhang (2010). "Numbers of genes in the NBS and RLK families vary by more than four-fold within a plant species and are regulated by multiple factors". In: *Nucleic Acids Research* 38.19.

Zhang, Yao, Yuancong Wang, Jingyan Liu, Yanglin Ding, Shanshan Wang, Xiaoyan Zhang, Yule Liu, and Shuhua Yang (2017). "Temperature-dependent autoimmunity mediated by chs1 requires its neighboring TNL gene SOC3". In: *New Phytol.* 213.3.

3.7. Online Methods

We characterized NLR gene variation in *A. thaliana*. NLR'omes were generated for a diverse set of accessions by targeted sequencing of long NLR containing fragments (SMRT RenSeq). R gene enrichment sequencing (RenSeq) is a targeted enrichment strategy that uses synthetic biotinylated RNA probes to capture DNA based on similarity (Jupe et al. 2013). Witek and collaborators (Witek et al. 2016) combined RenSeq with PacBio (SMRT RenSeq), to obtain long and curated reads that were used to unambiguously define NLR clusters, map and clone novel resistance genes (Witek et al. 2016). This method allows the specific sequencing and assembly of an organism's NLR gene complement, and several kilobases of flanking DNA sequences.

NLR genes were assembled, annotated, and classified by domain content. The core- and pan-NLR'ome was defined using orthogroups, and variation saturation was shown. Haplotype saturation was shown, and the selective forces that act on specific NLRs, domains, and positions were quantified. Evolutionary coupled co-evolving NLRs were detected.

3.7.1. NLR'ome Generation

3.7.1.1. Accession Selection

A. thaliana accessions were selected to attempt to cover the species' NLR gene diversity. For that, we sequenced the NLR'ome of 65 accessions that include a subset of 20 naturally occurring diverse accessions known as 'relicts' (mostly Iberian accessions that contain an unusually high amount of genetic diversity) (1001_Genomes_Consortium 2016), a subset of 19 in-lab selected phenotypically diverse panel known as MAGIC founders (Kover et al. 2009; Scarcelli et al. 2007), and a set of remaining accessions with high diversity on whole genome level and representing the different known haplotypes (1001_Genomes_Consortium 2016) (table 3.B.2).

3.7.1.2. Accession Verification

Routine seed stock genotyping prevents sample contamination (Pisupati et al. 2017). At a late stage of this project, 46 accessions were re-sequenced as part of a routine seed stock verification effort and the accessions were determined using `SNPmatch` as described (Pisupati et al. 2017). Three mis-labeled accessions were found in our dataset (Identifiers 7063, 9911, 9658). Their accession names and accession IDs from the 1001 genomes project were corrected and reported. For the sake of contiguity, their identifiers were not changed.

3.7.1.3. SMRT RenSeq

Genomic libraries were enriched for NLRs and sequenced using PacBio long read technology. Library preparations were performed collaboratively in three labs (UNC, MPI, and TSL) with only minor handling differences. DNA was extracted and fragmented to

2-5kb pieces for long read circular consensus sequencing (CCS). It was prepared using either the DNeasy plant Maxi kit (UNC), a custom high molecular weight DNA extraction protocol (MPI), or grinding in Shorty buffer (20% 1M Tris HCl pH 9, 20% 2M LiCl, 5% 0.5M EDTA, 10% SDS, 45% dH2O), followed by phenol chloroform extraction and precipitation with isopropanol (TSL). DNA was fragmented into shorter pieces using either Covaris Red miniTubes (Intensity=1, DutyCycle=20%, Cycles per Burst=1000, Treatment time=600s, Temperature=20°C, Water level=15, Sample volume=200µl) or Covaris g-tubes using manufacturer instructions for a targeted size of 6kb. The DNA was purified using 0.4x AMPure XP beads according to the manufacturer's instructions.

Libraries for NLR enrichment were constructed using the 'NEBNext Ultra DNA Library Prep Kit for Illumina'. Sixteen accessions from TSL were prepared for multiplexed sequencing, by introducing custom barcoded adapters (dual 8 bp index) instead of the standard ones (table 3.B.1). For the PCR amplification, 5-10µl adaptor ligated DNA was used together with 25µl 2xKAPA HiFi HotStart ReadyMix, 1µl Index and Universal PCR Primer, and 13-18µl water (to a total volume of 50µl). Initial Denaturation (94°C for 4min) was followed by at least 8 cycles (denaturation: 94°C for 30sec, annealing: 65°C for 30sec, extension: 68°C for 4min) and a final extension (68°C for 10min).

The genomic libraries were enriched for NLR genes. Roughly 1.4 Mb of the reference Col-0 *A. thaliana* genome contains NLRs. Baits were designed to hybridize with NLR containing genomic DNA regions, and only bound fragments were sequenced. 20,000 synthetic 120 nt biotinylated RNA probes (bait library), complementary to 736 known NLR genes from *A. thaliana* TAIR10 (Swarbreck et al. 2008), *Arabidopsis lyrata* (Hu et al. 2011), *Brassica rapa* (Wang et al. 2011), *Aethionema arabicum* (Haudry et al. 2013) and *Eutrema parv*µl*um* (Yang et al. 2013), were ordered as a MYbaits kit (Mycroarray, now Arbor Biosciences, USA). For select *Arabidopsis* genes additional alleles were included, as well as manually selected non-repetitive intron regions to improve capture in genes with introns longer than 350 bp. 100-500 ng of the libraries were hybridized with the baits using half of the reaction volume suggested in MYbaits v3.0 protocol with the following modifications: For each capture reaction, hybridization mix was prepared using 10µl Hyb#1, 4µl Hyb#3, 0.4µl Hyb#2 and 0.4µl Hyb#4; library mix with 2.5µl SeqCAP (Roche), 0.3µl Block#3 and 3µl gDNA Library; capture mix with 2.5µl Bait library and 0.5µl RNase block. Following the manufacturer's cycling conditions, we brought the mixes to a hybridization temperature of 65°C and transferred 5µl of the library mix and 5.5µl of the hybridization mix to the baits. After 16 to 24 hours hybridization the enriched libraries were recovered using 50µl Dynabeads MyOne Streptavidin C1 beads (Invitrogen). Binding and washing was carried out according to Mybaits 3.0 manual without the use of Hyb#4. Incubation of the captured libraries with the streptavidin beads was increased to 45 minutes. 30µl molecular biology grade water was used to re-suspend the DNA. The captured libraries were amplified for 18-30 cycles using the KAPA HiFi DNA Polymerase and the protocol for cycling conditions given in the previous paragraph.

Libraries were prepared for long read sequencing. PacBio libraries for MPI data were prepared using the '2 kb Template Preparation and Sequencing' protocol, and were size selected for 2-5kb using a BluePippin (0.75% Agarose Dye-Free/0.75% DF 2-6kb Marker

S1, Start=2000, End=2000). PacBio library prep for UNC data was done using the manufacturer's recommended procedure for 5kb template preparation and sequencing, and size selection for fragments over 3kb was done using a SAGE-ELF apparatus (SAGE Science) using 0.75% gel cassettes, size-based separation mode, target value 3kb and target well 10. All wells containing fractions above 3kb were pooled. Libraries for TSL data were prepared by size selecting fragments >3kb from the captured library using a SAGE-ELF apparatus (SAGE Science) as described above.

Quality control of all libraries was performed using Qubit (Invitrogen) and Bioanalyzer (Agilent). The PacBio RS II sequencing platform and P6-C4 chemistry was used to sequence each accession or multiplexed pool on individual SMRT cells. Sequencing of several accessions was repeated in order to obtain sufficient output reads (table 3.B.3).

3.7.1.4. Read Correction

Raw reads were used to produce highly accurate corrected reads. Circular consensus sequencing produced overlapping raw reads that were self-corrected to consensus reads which reduces the read error from 17% to 2% (CCS; version 2.0.0; defaults; Travers et al. (2010). Where indexing was employed, corrected sequences were de-multiplexed using a custom script. One combined CCS read dataset was created for accessions that were sequenced on more than one SMRT cell. Only CCS reads with more than 99% per base accuracy were considered further (fig. 3.A.13a).

3.7.1.5. Assembly

Reads were assembled to rebuild NLR containing regions (`Canu`; version 1.3; -pacbio-corrected, trimReadsCoverage=2, errorRate=0.01, genomeSize=2m; Koren et al. (2017)). Expected genome size was adjusted to 2Mb which reflects the proportion of the genome captured with RenSeq (1.4Mb NLR genes + flanking regions). Error rate reflected the input data quality after read correction. Read ends were trimmed using a minimum evidence of two reads. Contigs were removed if they were fully contained in a larger contig with >99.5% identity. The final assembly size and contig length distribution can be seen in fig. 3.A.13b.

3.7.1.6. Annotation

Coding and non-coding elements were annotated. Evidence and profile based methods were integrated in the `MAKER` pipeline (version 2.32; pred_flank=150, keep_preds=1, split_hit=3200, ep_score_limit=95, en_score_limit=95; Campbell et al. (2014)). Genes were predicted with `AUGUSTUS` (version 3.1.0; defaults; Stanke et al. (2004)) and `SNAP` (version 2006-07-28; defaults; (Korf 2004)). `AUGUSTUS` used the default 'arabidopsis' profile for gene prediction, and `SNAP` used a custom Hidden Markov Model (hmm) based on NB-ARC and/or TIR containing genes. Gene predictions were improved using Col-0 proteins and transcripts from Araport11 (Araport11_genes.20151202.pep.fasta, Araport11_genes.20151202.mRNA.fasta; *Araport Prerelease* (2018)).

Protein and transcript evidence was considered only if its mapping quality was high enough (see above for ep_score_limit and en_score_limit). Repeat-masked regions were not used for gene prediction (`RepeatMasker`; version open-4.0.5; model_org= arabidopsis; Smit et al. (2018)).

Capsella rubella and *Arabidopsis lyrata* reference annotations were revised to create reliable sets of NLRs for those outgroups. Reference annotations, evidence and gene predictions were integrated in `MAKER`. RNA-seq data guided gene prediction with `BRAKER1` (version 1.9; defaults; Hoff (2016)). Reads from silique, root, stem, leaf, and flower (PRJNA336053; PE; 100bp; 5-10MB; Wang et al. (2016)) were mapped to the reference genomes using `HISAT2` (version 2.0.5; –no-mixed –no-discordant; Kim et al. (2015)). Mapped reads guided gene prediction and were also used to assemble transcripts (`Cufflinks`; version 2.2.1; defaults; Trapnell et al. (2012)). Gene predictions were compared to reference gene annotations using `MAKER` (pred_gff, model_gff). Evidence mappings were used to choose the best annotation per locus. Reference genomes and annotations were taken from Phytozome (Phytozome 2018a,b). Assembled transcripts acted as the primary evidence (est_gff), re-annotated *A. thaliana* NLR transcripts and proteins were used as alternative evidence (altest, protein).

3.7.1.7. Web Apollo

Gene models were integrated into Web Apollo for manual inspection (version 2.0.4; http://ann-nblrrome.tuebingen.mpg.de/annotator/index; Lee et al. (2013)). Biological evidence lines were added to evaluate the quality of the gene models.

Transcripts and proteins from Col-0 showed the similarity of a gene model to reference gene annotations. Conserved genes were detected by mapping proteins and transcripts strictly with a minimal mapping score of 95%, duplicated and diversified genes were detected using a more relaxed minimal mapping score of 50% (`Exonerate`; version 2.2.0; Slater et al. (2005)). Pseudogenes were also detected using the relaxed mapping score.

Protein domains indicated conserved parts in genes. They were predicted for gene models and for `AUGUSTUS` gene predictions using Pfam hmms and coiled coils (`InterProScan`; version 5.20-59.0; -dp -iprlookup -appl Pfam,Coils; Zdobnov et al. (2001)). The predicted genes were added to see how well they agreed with the final gene models. Genes from `AUGUSTUS` and `SNAP` predictions were used.

Repeats often mark genomic regions with complicated annotations. `RepeatMasker` results were visualized. Diverged repeats in outgroups were additionally masked and visualized (repeat_protein=te_proteins.fasta provided by `MAKER`).

Raw reads showed the reliability of contig assemblies. Positions were more reliable if many reads could be mapped (`pbalign`; version 3.0; defaults; Tyagi et al. (2008)) and if many reads were used to construct the contig at that position (see section 3.7.1.5). RNA-seq transcripts and mapped reads indicated intron-exon boundaries.

3.7.1.8. Manual Reannotation

Genes containing NB-ARC or TIR domains were manually inspected to create accurate and reliable annotations (section 3.C.1). Gene models were evaluated using biological evidence in Web Apollo. Incorrectly fused genes were split, and incorrectly split genes were merged. Col-0 protein and transcript mappings were used to detect wrongly fused or split gene models. Genes were split if several proteins or transcripts mapped next to each other within one model. Genes were merged if protein or transcript mappings spanned several models. Both cases often showed disagreeing gene predictions. Additional features of fused genes were extremely long introns, or pseudogene mappings.

Gene structures were corrected. Intron-, exon-, and UTR boundaries were refined, and alternative splice forms were added. Evidence from protein and transcript mappings, as well as RNA-seq read mappings was considered. Genes were flagged with 'corbound' if exon-intron structures were changed without direct protein or transcript evidence, and 'cortrans' was used, if translation start points were changed. Exceptions were detected and evaluated. Non-canonical splice sites were confirmed using reference proteins and transcripts. Rare erroneous reference annotations were corrected using TAIR10 annotations (TAIR 2018). Genes were flagged with 'pseudogene' if a pseudogene from Araport11 was aligned to the same region.

Incomplete genes and incorrectly annotations were flagged. Genes at contig borders were flagged as 'truncated' if confirmed by protein or transcript mappings. Rarely, genes were extensively changed to rescue domain structures. These genes were flagged with 'mod'. Wrong gene models due to misassembled contigs were detected. Genes were flagged with misassembly' if base calls were contradicted reliably by CCS read mappings.

Manual re-annotation was necessary to secure the reliability of NLR gene models[1].

3.7.1.9. Paired NLRs

We generated a list of paired NLRs containing the nine Col-0 divergently transcribed TNLs sharing a genetic arrangement similar to the RPS4/RRS1 pair (Narusaka et al. 2009). We added seven additional divergently transcribed pairs identified by manual inspection of 138 Col-0 genes that contained a TIR domain. We also used a CNL clone list (Dangl lab, unpublished) to mine the Col-0 genome for consecutive genes and included six paired CNL-CNL loci, of which only two are divergently transcribed. During manual curation we further identified one divergently transcribed pair of TNLs with no Col-0 allele and included it, too.

Identification of sensor-executor pairs To further examine pair evolution, we made a narrower list of pairs. These are in head to head genetic orientation (in either the Col-0 reference genome, or in the assembled contigs where these gene pairs exist) and phylogenetically in the clades containing either RPS4 or SOC3 executor TNLs or the

[1]The final version of the manuscript will contain a link to the re-annotated gff files. Currently, this data is not publicly accessible.

clades containing RRS1 or CHS1 sensor TN(L)s. The NB domain alignment based phylogeny used to make this decision is presented in fig. 3.A.2.

There are 16 such pairs identified in the pan-NLR'ome, two of which do not appear in the Col-0 reference genome. As a control group to test the possibility that genetic proximity could lead to co-evolution or conservation of popμlation genetic characteristics, we identified a set of control pairs. These are pairs of NLR-encoding genes that are less than 4kb apart in the Col-0 reference genome, but are not classified as sensor/executor pairs. We identified 15 such pairs, a list of all pairs is provided in table 3.B.5.

3.7.1.10. Classification

Each gene that contained an NB, a TIR, an RPW8 domain, or a combination of those was defined as an NLR. The presence of LRR or CC motifs alone did not suffice our criteria. As a first subdivision (fig. 3.1c), we defined TNLs (at least a TIR domain), CNLs (CC+NB domain), RNLs (at least a RPW8 domain), and NLs (at least a NB domain).

The second subdivision defined 26 different groups by the different combinations of TIR, CC, NB, RPW8, LRR, and X (other Integrated Domains (ID)) independent of arrangement and number (fig. 3.A.1).

As mentioned earlier (see section 3.7.1.7), protein domains were predicted using Pfam hmms. Coiled-coil (CC) motifs were refined in NLR genes using a majority vote from different prediction programs. In order to secure the correct annotation, `Coils` (2.2.1; `InterProScan`-defaults; Lupas et al. (1991)), `Paircoil2` (defaults; McDonnell et al. (2006)), and the `NLR-parser` (v.2; defaults; Steuernagel et al. (2015)) predictions were compared to each other. `Coils` and `Paircoil2` both use databases of many known coiled-coils, whereas the `NLR-parser` uses two NLR-specific coiled-coil motifs (motif16 and motif17) (Steuernagel et al. 2015). CC signatures were considered credible if overlapping predictions existed in at least two of the three methods. Notably, CCs of functional NLRs previously published as CNLs are not always confirmed (table 3.B.6).

3.7.1.11. Architectures

An architecture is defined as the collapsed protein domain set in an NLR gene. Domains occurring multiple times are reported only once fig. 3.A.3b). The domain annotations, as well as the ordered and the collapsed architectures are available on github.[2]

A high order domain composition classification distinguishes between canonical and non-canonical domain architectures. Canonical architectures are strictly composed by any combination of NB (Pfam accession PF00931), TIR (PF01582), RPW8 (PF05659), LRR (PF00560, PF07725, PF13306, PF13855), or the Coiled-Coil structural motifs (fig. 3.A.3a)). Non-canonical architectures contain at least one ID, as defined in Baggs et al. (2017).

[2]The final version of the manuscript will contain a link to a github folder that contains this information. Currently, this folder is not publicly accessible.

To be able to compare the pan-NLR'ome to the reference Col-0 accession, we removed the Col-0 RenSeq dataset before performing the architecture analyses (section 3.4.3).

In order to identify novel and recurring domain arrangements, we compared the reference Araport11 Col-0 NLRs, with the pan-NLR'ome and the NLR'ome of 22 Brassicaceae species (table 3.B.7). In all domain architecture comparisons, we explicitly excluded the Col-0 RenSeq gene models to enable the comparison to the reference Col-0, and included the Brassicaceae sets, whenever required.

3.7.2. Figure Generation

All figure panels were generated using R (version 3.4.4; R Development Core Team (2008)) and RStudio (several versions; RStudio Team (2015)), if not stated otherwise. Used packages included ggplot2, rworldmap, yarrr (pirate plots), UpSetR (to visualize intersections), PerformanceAnalytics (Correlation plots), karyoploteR (OG-anchoring to Col-0 reference genome). Phylogenetic trees were visualized using iTOL (Letunic et al. 2016).

3.7.3. Pan-NLR'ome Generation

3.7.3.1. Generation

The pan-NLR'ome of A.thaliana was constructed using a protein-clustering approach. Each protein cluster contained a set of homology-related NLRs from different accessions, and was termed an 'orthogroup'. Singletons were proteins that did not cluster with any other protein. Protein clusters were generated with a three step procedure. All-against-all protein alignments were produced (`DIAMOND`; version 0.9.1.102; –max-target-seqs 13169 –more-sensitive –comp-based-stats; Buchfink et al. (2014)). Putative orthology and inparalogy relationships were identified (`orthAgogue`, commit 82dcb7aeb67c, –use_scores –strict_coorthologs; Ekseth et al. (2014)). Protein clusters were formed based on the orthology information (`mcl`; version 12-135; -I 1.5; Enright (2002)).

3.7.3.2. Refinement

The initial set of orthogroups was inspected for over-clustering by screening for paralogs within orthogroups. A protein alignment for each orthogroup with more than 4 members was generated (`T-Coffee`; version 11.00.8cbe486; mode mcoffee; Neelabh et al. (2016)). Protein sequence alignments were converted into the corresponding codon alignments (`PAL2NAL`; version 14, defaults; Notredame et al. (2000)). The resulting codon alignment was used to remove three different types of outliers, namely non-homologous, partly mistranslated and low similarity sequences (`OD-seq`; version 1.0; –analysis bootstrap; Jehl et al. (2015)). The remaining core sequences for each outgroup were again aligned in protein space. The protein alignments were converted into the corresponding codon alignments and used to infer a phylogenetic tree (`FastME`; version 2.1.5.1; -s -n -b 100; Lefort et al. (2015)). Each tree was used to detect simple paralogs (duplications in terminal branches) and complex paralogs (duplications spread across the whole

phylogeny). For orthogroups where less than or equal to 5% of its accession members showed duplications, all paralogs were removed. Otherwise the tree was split at (accession) duplication events (`ete3`; version 3.0.0b36; Huerta-Cepas et al. (2016)) and new orthogroups were created from the leaves of all resulting sub trees. Protein, codon alignments and trees were re-computed as stated above.

3.7.3.3. Annotation

The final set of refined orthogroups was annotated with metadata derived from transcript-based majority votes (e.g., classes), transcript-based counts (e.g., members with IDs, members flagged as paired, members flagged as clustered) or orthogroup-based counts and analysis (e.g., type, diversity statistics, positive selection, average tree branch length). Refined orthogroups were classified into three size-based categories after visual inspection of the orthogroup size density distribution fig. 3.3. Orthogroups with less than 13 members were typed as 'cloud', orthogroups with at least 52 members 'core', and those in between were typed 'shell'.

We classified orthogroups using protein domain architectures. We assigned a class and subclass to each orthogroup by using the majority vote from its members' architectures.

Diversity and neutrality statistics were calculated for each codon alignment of the refined orthogroups using `PopGenome` (version 2.2.4; Pfeifer et al. (2014)). The average tree-derived branch length for an orthogroup was defined as the sum of all branch lengths normalized by the orthogroup size. Positive selection tests were carried out using `HyPhy` (version 2.3.13; Pond et al. (2005)) using codon alignments and corresponding trees. Pervasive diversifying positive selection was detected with `FUBAR` (version 2.1; default parameters, Murrell et al. (2013)) and sites considered with a posterior probability $>=$ 0.95. Episodic diversifying positive selection was detected with `MEME` (version 2.0.1; default parameters; Murrell et al. (2012)) and sites considered with a p-value threshold $<=$ 0.01. An average expression percentage was estimated for each orthogroup using RNA-Seq data from the 1001 Genomes collection (Kawakatsu et al. 2016). For each accession, a pseudo-transcriptome was generated from the accession-specific NLR transcripts and all non-NLR transcripts from the reference Col-0. The NLR genes' introns were added to the pseudo-transcriptome for expression filtering. Transcript abundance was quantified with pseudoalignments of RNA-Seq reads from 727 accessions of the 1001 Genomes collection (`kallisto`, v.0.43.0, –single -l 200 -s 25 -b 100 –bias; Bray et al. (2016)). The data was further processed with `R` (v.3.4.1). Abundance was normalized (`DESeq2`; v.1.16.1; estimateSizeFactor; Love et al. (2014)) and expressed NLRs were defined using a per-accession expression threshold. Expression counts from introns were used to compute a background expression density distribution and subtracted from the density distribution of all NLR expression counts. The lowest expression level with a density >0 was used as minimum expression threshold. On average, NLRs were considered expressed with an expected count $>=$ 175. Finally, for each NLR, the percentage of accessions that provided reliable expression was calculated. Furthermore, we consulted the AtGenExpress expression atlas to gauge absolute expression level, bias in leaf vs root specificity of expression and the pathogen inducibility of Col-0 NLRs. NLR genes

were broadly divided into low, medium and high expression groups, based on whether at least two samples had absolute signal values in the developmental data sets that were $20 <$ expression < 100, $100 <$ expression $< 1{,}000$, $1{,}000 <$ expression. Genes that had generally absolute signal values below 20 were characterized as marginally expressed. If average expression in leaf and rosette samples was at least twice of that in root samples, or vice versa, genes were considered tissue biased in expression. Note that differences between tissues can be much larger, exceeding 100 fold. Pathogen inducibility was assessed from the AtGenExpress pathogen data set, based on consistent induction by at least two pathogen-related stimuli. The final Col-0 NLR annotation was amended to the respective orthogroups.

3.7.3.4. Visualization

Raw orthogroups and corresponding metadata were integrated in iTOL (Letunic et al. 2016) for visualization and reinspection (https://itol.embl.de/shared_projects.cgi, iTOLlogin: fbemm).

Protein trees showed the evolutionary history of each orthogroup. Similarity between members was also reflected in branch lengths and bootstrap values. The multiple sequence alignment showed sequence variation on base pair level. The identifiers of refined orthogroups were added to show over-clustered orthogroups and outliers. The domain architecture and the protein length was plotted to compare orthogroup members structurally. Transposable elements (TEs) are known to influence gene activity or be the cause of gene duplication. TEs in exons, introns, and 2kb up- or downstream of NLRs were integrated in iTOL.

Sub-clustering might be related to accession-based metadata, thus we included for each protein if its accession belonged to the relict group, the geographic origin, and the admixture group.

A few misannotated genes (table 3.B.4) were detected in OGs and removed from the NLR'ome.

3.7.4. Saturation Analysis

Orthogroup stability determined the saturation of our dataset. Accessions were randomly removed and the number of remaining orthogroups served as a proxy for saturation. Accessions were removed in steps of one, and each step was repeated 10000 times.

3.7.5. Assembly Quality

3.7.5.1. Quality Scores

We used pseudo-heterozygous SNP calls created by mis-mapped reads, as a measurement for assembly quality. A read that couldn't be mapped to its correct NLR origin because

the NLR was not assembled, instead mapped to a similar NLR and created pseudo-heterozygous SNPs. The quality was calculated from the ratio of pseudo-heterozygous SNPs and the total amount of mapped bases.

The pseudo-heterozygosity was determined using the corrected CCS reads. A pseudo-genome was constructed for each RenSeq assembly by combining the assembled contigs with chromosomes from the tair10 reference. To avoid mis-mappings, non-NLRs were masked on the RenSeq contigs, and NLRs were masked on the reference chromosomes (fig. 3.A.3c)). CCS reads were then mapped to those pseudo-genomes (`minimap2`; 2.9-r748-dirty; -x map-pb; Li (2018b)). SNPs were called for NLR genes using high quality mappings only (`htsbox` pileup; r345; -S250 -q20 -Q3 -s5; Li (n.d.)).

The number of pseudo-heterozygous sites (hetsites) was compared to the total number of mappable NLR gene bases (totalsites).The quality was calculated as logarithmically linked to the ratio of pseudo-heterozygous calls to the total amount of mapped bases (fig. 3.A.13d).

3.7.5.2. Completeness Assessment

The completeness of an accession's NLR complement was derived from quality and completeness relationships in the reference Col-0. We created the correlation between completeness and quality for Col-0 sub-assemblies using different amounts of input data. The corrected CCS reads from Col-0 were sub-sampled from 100% to 1% in 1% steps (`seqtk` sample; v.1.0-r82-dirty; defaults; Li (2018a)). 100% of the data correspond to 26 639 reads with a N50 read length of 2846 bases and 77.98 Mb sequence in total. The sub-sampled datasets were assembled with `Canu`. All genes from the original RenSeq Col-0 assembly were mapped to each sub-assembly to detect assembled NLRs. NLR transcripts were extracted using those alignments (`exonerate`; v.2.2.0; –model est2genome –bestn 1 –refine region –maxintron 546; Slater et al. (2005)). The quality of each sub-assembly was assessed as described above (fig. 3.A.13c). The completeness of a sub-assembly was determined as the fraction of the full reference Col-0 NLR complement that was assembled. NLR transcripts were evaluated using `rnaQUAST` (version 1.5.0; defaults; Bushmanova et al. (2016)) with the tair10 reference genome and Araport11 NLRs. The completeness was calculated by dividing the amount of covered NLR genes (in bases) by the total length of the Araport11 NLRs (fig. 3.A.13c). The relation between completeness and quality of the tested Col-0 sub-assemblies was used to infer completeness values for the other accessions. Each accession's quality was used to find the corresponding completeness value from the tested Col-0 sub-assemblies (fig. 3.A.13d).

3.7.5.3. Similarity to Col-0

We determined if the similarity of an accession to the reference Col-0 influenced its quality (fig. 3.A.13e). RenSeq assemblies were mapped against the Col-0 assembly (`minimap2`; 2.9-r748-dirty; defaults; Li (2018b)) and SNPs were called in NLR gene regions (`htsbox` pileup; r345; defaults; Li (n.d.)). Only biallelic SNPs were used to calculate the Identity By State (IBS) value for each accession compared to Col-0 (`SNPRelate`_1.10.2;

method='biallelic'; Zheng et al. (2012)).

3.7.5.4. Orthogroup co-occurrence / Anchoring analysis

Annotated non-NLR genes in the 65 accessions were clustered into orthogroups. For each NLR in an orthogroup, we tested which other NLR or non-NLR orthogroups co-occurred on the same assembled contig. Co-occurrence matrices were used to calculate bidirectional matrices showing orthogroups as nodes, and scaled to the number times an OG-OG co-occurrence was detected (edges). We also obtained per accession OG co-occurrence matrices to visualize in an UpSet plot the most common OG-OG arrangements observed across contigs. Networks were visualized in Cytoscape v.3.5.1 (Shannon et al. 2003), running on Java v. 1.8.0_151. Putative paired NLRs were identified by testing OG enrichment in annotation flags. Enrichment was done by Fisher and hypergeometric tests and FDR. All enrichments with a q-value below 0.1 were significant.

References

1001_Genomes_Consortium (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". In: *Cell* 166.2. URL: http://dx.doi.org/10.1016/j.cell.2016.05.063.

Araport Prerelease (2018). URL: https://araport-dev.tacc.utexas.edu/downloads/Araport11%7B%5C_%7DPreRelease%7B%5C_%7D20151202 (visited on 10/15/2018).

Baggs, Erin, G Dagdas, and Ksenia V. Krasileva (2017). "NLR diversity, helpers and integrated domains: making sense of the NLR IDentity". In: *Current Opinion in Plant Biology* 38.Figure 1. URL: http://linkinghub.elsevier.com/retrieve/pii/S1369526616301741.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter (Aug. 2016). "Erratum: Near-optimal probabilistic RNA-seq quantification". In: *Nat. Biotechnol.* 34.8.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson (2014). "Fast and sensitive protein alignment using DIAMOND". In: *Nature Methods* 12.1.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, Vladimir Suvorov, and Andrey D. Prjibelski (2016). "RnaQUAST: A quality assessment tool for de novo transcriptome assemblies". In: *Bioinformatics* 32.14.

Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell (2014). "Genome Annotation and Curation Using MAKER and MAKER-P". In: *Current Protocols in Bioinformatics*. arXiv: NIHMS150003.

Ekseth, Ole Kristian, Martin Kuiper, and Vladimir Mironov (2014). "OrthAgogue: An agile tool for the rapid prediction of orthology relations". In: *Bioinformatics* 30.5.

Enright, A. J. (2002). "An efficient algorithm for large-scale detection of protein families". In: *Nucleic Acids Research* 30.7. arXiv: journal.pone.0035671 [10.1371]. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/30.7.1575.

Haudry, Annabelle, Adrian E. Platts, Emilio Vello, Douglas R. Hoen, Mickael Leclercq, Robert J. Williamson, Ewa Forczek, Zoé Joly-Lopez, Joshua G. Steffen, Khaled M. Hazzouri, Ken Dewar, John R. Stinchcombe, Daniel J. Schoen, Xiaowu Wang, Jeremy Schmutz, Christopher D. Town, Patrick P. Edger, J. Chris Pires, Karen S. Schumaker, David E. Jarvis, Terezie Mandáková, Martin A. Lysak, Erik Van Den Bergh, M. Eric Schranz, Paul M. Harrison, Alan M. Moses, Thomas E. Bureau, Stephen I. Wright, and Mathieu Blanchette (2013). "An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions". In: *Nature Genetics* 45.8. URL: http://dx.doi.org/10.1038/ng.2684.

Hoff (2016). "BRAKER1". In: *Bioinformatics* 32.5.

Hu, Tina T., Pedro Pattyn, Erica G. Bakker, Jun Cao, Jan Fang Cheng, Richard M. Clark, Noah Fahlgren, Jeffrey A. Fawcett, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Jesse D. Hollister, Stephan Ossowski, Robert P. Ottilar, Asaf A. Salamov, Korbinian Schneeberger, Manuel Spannagl, Xi Wang, Liang Yang, Mikhail E. Nasrallah, Joy Bergelson, James C. Carrington, Brandon S. Gaut, Jeremy Schmutz, Klaus F.X. Mayer, Yves Van De Peer, Igor V. Grigoriev, Magnus Nordborg, Detlef Weigel, and Ya Long Guo (2011). "The Arabidopsis lyrata genome sequence and the basis of rapid genome size change". In: *Nature Genetics* 43.5.

Huerta-Cepas, Jaime, François Serra, and Peer Bork (2016). "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". In: *Mol. Biol. Evol.* 33.6.

Jehl, Peter, Fabian Sievers, and Desmond G Higgins (Aug. 2015). "OD-seq: outlier detection in multiple sequence alignments". In: *BMC Bioinformatics* 16.

Jupe, Florian, Kamil Witek, Walter Verweij, Jadwiga Sliwka, Leighton Pritchard, Graham J Etherington, Dan Maclean, Peter J Cock, Richard M Leggett, Glenn J Bryan, Linda Cardle, Ingo Hein, and Jonathan D G Jones (Nov. 2013). "Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations." In: *The Plant journal : for cell and molecular biology* 76.3. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3935411%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Kawakatsu, Taiji, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, Joseph R Nery, Cesar Barragan, Yupeng He, Huaming Chen, Manu Dubin, Cheng Ruei Lee, Congmao Wang, Felix Bemm, Claude Becker, Ryan O'Neil, Ronan C O'Malley, Danjuma X Quarless, Carlos Alonso-Blanco, Jorge Andrade, Felix Bemm, Joy Bergelson, Karsten Borgwardt, Eunyoung Chae, Todd Dezwaan, Wei Ding, Joseph R Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G Grimm, Angela Hancock, Stefan R Henz, Svante

Holm, Matthew Horton, Mike Jarsulic, Randall A Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Chen Ruei Lee, Dazhe Meng, Todd P Michael, Richard Mott, Ni Wayan W. Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Novikova, F. Xavier Pico, Alexander Platzer, Fernando A Rabanal, Alex Rodriguez, Beth A Rowan, Patrice A. Salome, Karl Schmid, Robert J Schmitz, Ü Seren, Felice Gianluca G. Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, Xuefeng Zhou, Nicholas J. Schork, Detlef Weigel, Magnus Nordborg, and Joseph R. Ecker (2016). "Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions". In: *Cell* 166.2.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg (2015). "HISAT: A fast spliced aligner with low memory requirements". In: *Nature Methods* 12.4. arXiv: 15334406.

Koren, Sergey, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy (2017). "Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". In: *Genome research*. arXiv: 071282.

Korf, Ian (2004). "Gene finding in novel genomes". In: *BMC Bioinformatics* 5.59.

Kover, Paula X., William Valdar, Joseph Trakalo, Nora Scarcelli, Ian M. Ehrenreich, Michael D. Purugganan, Caroline Durrant, and Richard Mott (2009). "A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana". In: *PLoS Genetics* 5.7. arXiv: 15334406.

Lee, Eduardo, Gregg a Helt, Justin T Reese, Monica C Munoz-Torres, Chris P Childers, Robert M Buels, Lincoln Stein, Ian H Holmes, Christine G Elsik, and Suzanna E Lewis (2013). "Web Apollo: a web-based genomic annotation editing platform." In: *Genome biology* 14.8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0082%20http://www.ncbi.nlm.nih.gov/pubmed/24000942.

Lefort, Vincent, Richard Desper, and Olivier Gascuel (2015). "FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program". In: *Mol. Biol. Evol.* 32.10.

Letunic, Ivica and Peer Bork (2016). "Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees". In: *Nucleic acids research* 44.W1.

Li, Heng (n.d.). *htsbox*. URL: https://github.com/lh3/htsbox.

– (2018a). *seqtk Toolkit for processing sequences in FASTA/Q formats*. URL: https://github.com/lh3/seqtk (visited on 10/15/2018).

– (2018b). "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics*. arXiv: 1708.01492. URL: http://arxiv.org/abs/1708.01492.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol.* 15.12.

Lupas, Andrei, M van Dyke, and Jeff Stock (1991). "Predicting coiled coils from proteins sequences". In: *Science* 252.

McDonnell, A. V., T. Jiang, A. E. Keating, and Bab Berger (2006). "Paircoil2: Improved prediction of coiled coils from sequence". In: *Bioinformatics* 22.3.

Murrell, Ben, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Sergei L Kosakovsky Pond, and Konrad Scheffler (2013). "FUBAR: a fast, unconstrained bayesian approximation for inferring selection". In: *Mol. Biol. Evol.* 30.5.

Murrell, Ben, Joel O Wertheim, Sasha Moola, Thomas Weighill, Konrad Scheffler, and Sergei L Kosakovsky Pond (2012). "Detecting individual sites subject to episodic diversifying selection". In: *PLoS Genet.* 8.7.

Narusaka, Mari, Ken Shirasu, Yoshiteru Noutoshi, Yasuyuki Kubo, Tomonori Shiraishi, Masaki Iwabuchi, and Yoshihiro Narusaka (2009). "RRS1 and RPS4 provide a dual Resistance-gene system against fungal and bacterial pathogens". In: *Plant Journal* 60.2.

Neelabh, Karuna Singh, and Jyoti Rani (2016). "Sequential and Structural Aspects of Antifungal Peptides from Animals, Bacteria and Fungi Based on Bioinformatics Tools". In: *Probiotics Antimicrob. Proteins* 8.2.

Notredame, C, D G Higgins, and J Heringa (Sept. 2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment". In: *J. Mol. Biol.* 302.1.

Pfeifer, Bastian, Ulrich Wittelsbürger, Sebastian E Ramos-Onsins, and Martin J Lercher (2014). "PopGenome: an efficient Swiss army knife for population genomic analyses in R". In: *Mol. Biol. Evol.* 31.7.

Phytozome (2018a). *Arabidopsis lyrata v2.1*. URL: `https://phytozome.jgi.doe.gov/pz/portal.html%7B%5C#%7D!info?alias=Org%7B%5C_%7DAlyrata` (visited on 11/14/2018).

– (2018b). *Capsella rubella v1.0*. URL: `https://phytozome.jgi.doe.gov/pz/portal.html%7B%5C#%7D!info?alias=Org%7B%5C_%7DCrubella` (visited on 11/14/2018).

Pisupati, Rahul, Ilka Reichardt, Pamela Korte, Viktoria Nizhynska, Envel Kerdaffrec, Kristina Uzunova, Fernando Rabanal, Daniele Filiault, and Magnus Nordborg (2017). "Verification of Arabidopsis stock collections using SNPmatch - an algorithm for genotyping high-plexed samples". In: *bioarxiv*.

Pond, Sergei L Kosakovsky, Simon D W Frost, and Spencer V Muse (2005). "HyPhy: hypothesis testing using phylogenies". In: *Bioinformatics* 21.5.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.r-project.org.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc. Boston, MA. URL: http://www.rstudio.com/.

Scarcelli, N., J. M. Cheverud, B. A. Schaal, and P. X. Kover (2007). "Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus". In: *Proceedings of the National Academy of Sciences* 104.43. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0708209104.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker (Nov. 2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome Res.* 13.11.

Slater, Guy St C. and Ewan Birney (2005). "Automated generation of heuristics for biological sequence comparison". In: *BMC Bioinformatics* 6.

Smit, AFA, R Hubley, and P Green (2018). *RepeatMasker Open-4.0*. URL: http://www.repeatmasker.org (visited on 11/13/2018).

Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern (July 2004). "AUGUSTUS: a web server for gene finding in eukaryotes." In: *Nucleic acids research* 32.Web Server issue. URL: http://nar.oxfordjournals.org/content/32/suppl%7B%5C_%7D2/W309.full.

Steuernagel, Burkhard, Florian Jupe, Kamil Witek, Jonathan D G Jones, and Brande B H Wulff (Jan. 2015). "NLR-parser: Rapid annotation of plant NLR complements." In: *Bioinformatics (Oxford, England)*. URL: http://www.ncbi.nlm.nih.gov/pubmed/25586514.

Swarbreck, David, Christopher Wilks, Philippe Lamesch, Tanya Z. Berardini, Margarita Garcia-Hernandez, Hartmut Foerster, Donghui Li, Tom Meyer, Robert Muller, Larry Ploetz, Amie Radenbaugh, Shanker Singh, Vanessa Swing, Christophe Tissier, Peifen Zhang, and Eva Huala (2008). "The Arabidopsis Information Resource (TAIR): Gene structure and function annotation". In: *Nucleic Acids Research* 36.SUPPL. 1.

TAIR (2018). *TAIR10 genome release*. URL: https://www.arabidopsis.org/download/index-auto.jsp?dir=%7B%5C%%7D2Fdownload%7B%5C_%7Dfiles%7B%5C%%7D2FGenes%7B%5C%%7D2FTAIR10%7B%5C_%7Dgenome%7B%5C_%7Drelease (visited on 11/15/2018).

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nature Protocols* 7.3.

Travers, Kevin J., Chen Shan Chin, David R. Rank, John S. Eid, and Stephen W. Turner (2010). "A flexible and efficient template format for circular consensus sequencing and SNP detection". In: *Nucleic Acids Research* 38.15.

Tyagi, M, A G de Brevern, N Srinivasan, and B Offmann (2008). "Protein structure mining using a structural alphabet". In: *Proteins* 71.2.

Wang, Jinyan, Xilin Hou, and Xuedong Yang (2011). "Identification of conserved microRNAs and their targets in Chinese cabbage (Brassica rapa subsp. pekinensis)". In: *Genome* 54.12. URL: https://doi.org/10.1139/g11-069.

Wang, Jun, Feng Tao, Nicholas C Marowsky, and Chuanzhu Fan (Sept. 2016). "Evolutionary Fates and Dynamic Functionalization of Young Duplicate Genes in Arabidopsis Genomes". In: *Plant Physiology* 172.1. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5074645/.

Witek, Kamil, Florian Jupe, Agnieszka I Witek, David Baker, Matthew D Clark, and Jonathan D G Jones (Apr. 2016). "Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing". In: *Nature Biotechnology* advance on. URL: http://dx.doi.org/10.1038/nbt.3540.

Yang, Ruolin, David E. Jarvis, Hao Chen, Mark A. Beilstein, Jane Grimwood, Jerry Jenkins, ShengQiang Shu, Simon Prochnik, Mingming Xin, Chuang Ma, Jeremy Schmutz, Rod A. Wing, Thomas Mitchell-Olds, Karen S. Schumaker, and Xiangfeng Wang (2013). "The Reference Genome of the Halophytic Plant Eutrema salsugineum". In: *Frontiers in Plant Science* 4.March. URL: http://journal.frontiersin.org/article/10.3389/fpls.2013.00046/abstract.

Zdobnov, E M and R Apweiler (Sept. 2001). "InterProScan–an integration platform for the signature-recognition methods in InterPro". In: *Bioinformatics* 17.9.

Zheng, Xiuwen, David Levine, Jess Shen, Stephanie M Gogarten, Cathy Laurie, and Bruce S Weir (2012). "A high-performance computing toolset for relatedness and principal component analysis of SNP data". In: *Bioinformatics* 28.24.

Appendix

3.A. Supplemental Figures

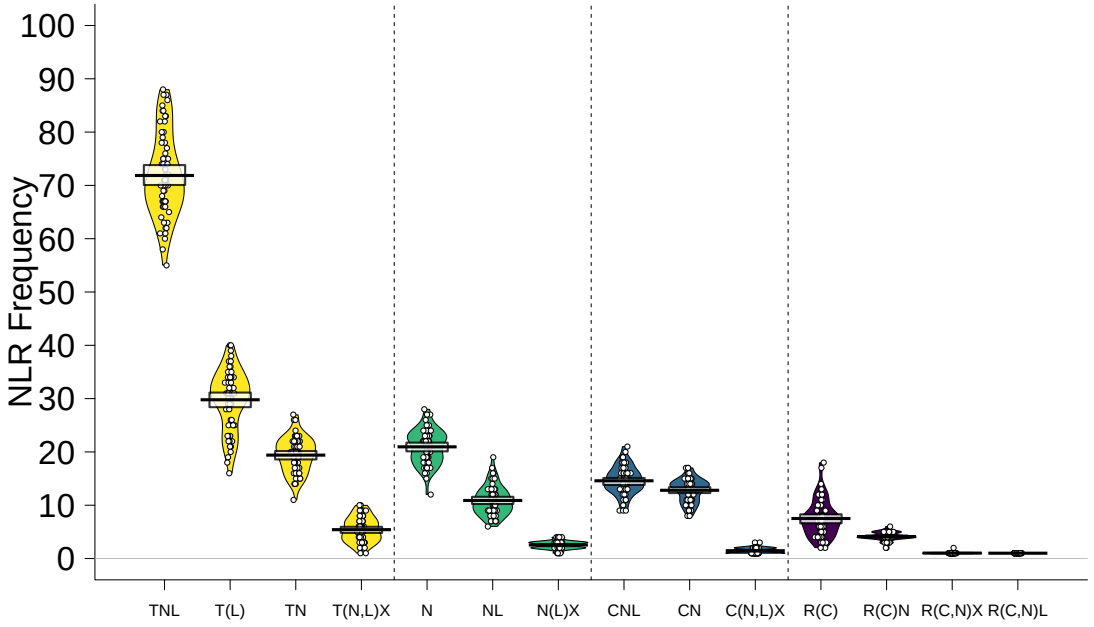


Figure 3.A.1.: For each subclass, the corresponding class is color coded (TNLs: yellow, NLs: green, CNLs: blue, and RNLs: purple), and classes are in addition divided by the vertical dashed lines. NLRs are grouped into subclasses by their domains content: T (TIR), N (NB), C (CC), R (RPW8), and X (all other integrated domains). Each domain must be present at least once, domains in brackets may be present. Domain order is not considered. The mean is shown as a solid black horizontal line and the 95 % Highest density Intervals (HDI: points in the interval have a higher probability than points outside) are shown as solid bands around the sample mean. All raw data points are plotted as open circles and the full densities are shown as a bean plot.

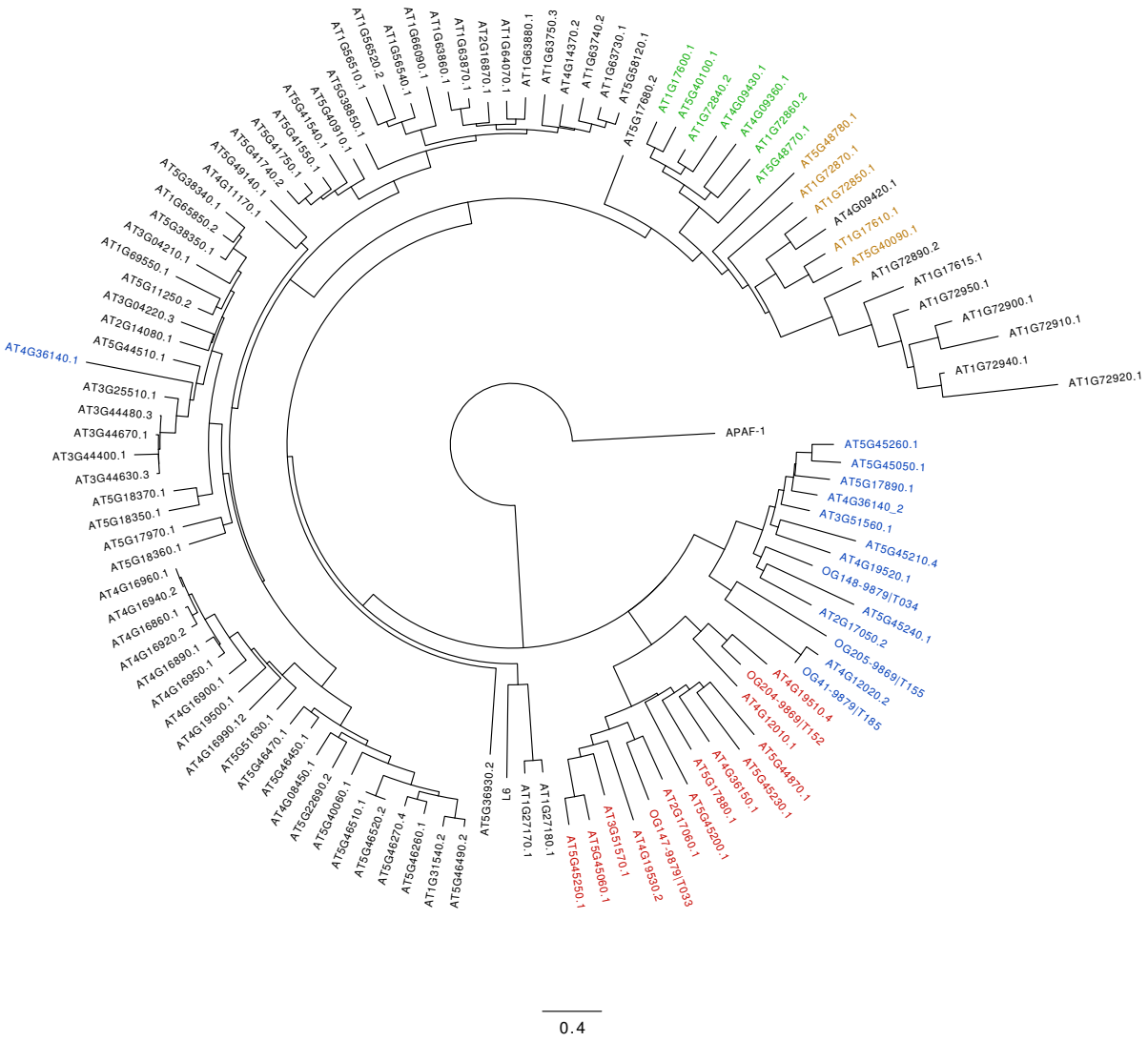


Figure 3.A.2.: Phylogenetic tree of TIR and NB containing proteins
Phylogenetic tree of TIR and NB containing proteins to discriminate sensors and executors in pairs. RPS4/RRS1-like and SOC3/CHS1-like paired TNLs fall into distinct subclades. These are indicated by color: RPS4-like (Red, executors), RRS1-like (Blue, sensors), SOC3-like (Green, executors) and CHS1-like (Yellow, sensors). This phylogeny was constructed by aligning the NB domain (240 amino acids) of all TIR and NB containing Col-0 proteins and selected additional representatives of pair flagged orthogroups (OGs) from the pan-NLR'ome that are not represented in Col-0 (identified by their OG and protein numbers). NB domains from APAF-1 (Human) and L6 (Flax) were also included. Amino acid sequences were aligned with `MUSCLE` (Neighbor joining clustering), refined by manual trimming and the phylogeny produced with the WAG+F maximum likelihood method allowing for 5 discrete Gamma categories. AT4G36140 contains two distinct NB domains, both of which were included and the second of which clusters with other RRS1-like NB domains. Scale bar represents amino acid substitutions per site.

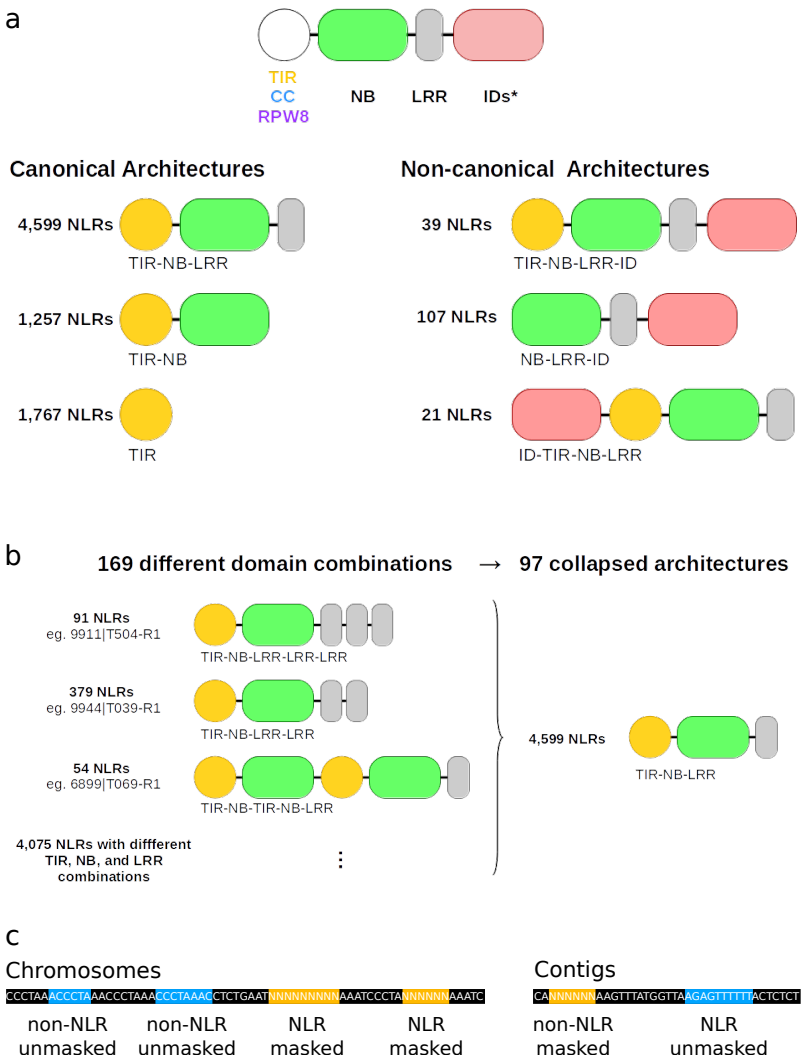


Figure 3.A.3.: Architectures and Pseudo-genomes
a) Examples of NLR domain architecture diversity. On top, a generic NLR, with an ID (Integrated Domain) shown at the C-terminus. IDs can also be found at the N-terminus, and more rarely between the three canonical domain types. b) Reduction of domain combinations by collapsing duplicated/repetitive domains. The number of NLRs grouped by each of the original architectures is shown on the left, along with one example that can be visualized in the genome browser (to be released soon). Ellipsis in the bottom left represent 19 other architectures containing 4,079 proteins exclusively composed of TIR, NB and LRR domains. The same strategy was applied to all other architectures containing at least one duplicated domain in the RPW8, NB and CC classes. c) Pseudo-genome generation. For each accession the pseudo-genome contains the assembled contigs with non-NLRs being masked, and the reference chromosomes of the tair10 reference with masked NLRs.

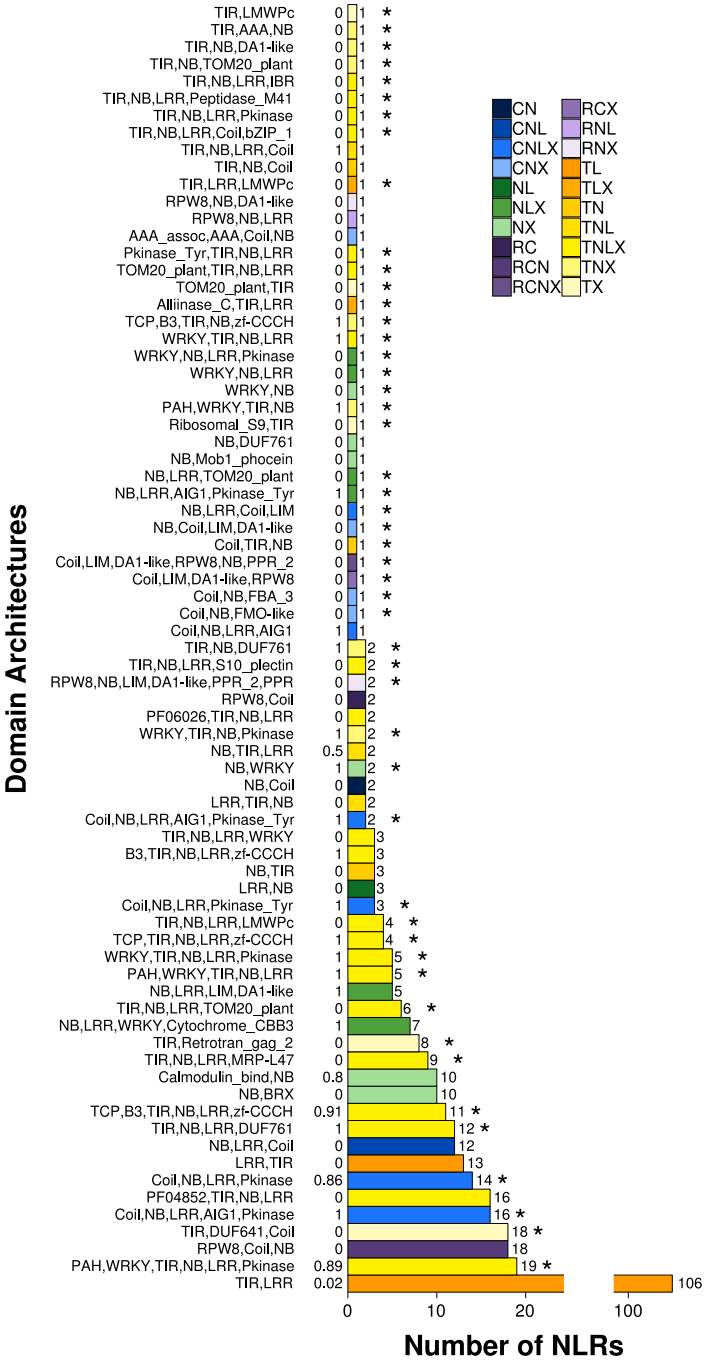


Figure 3.A.4.: Novel *A. thaliana* NLR architectures
Full set of the novel *A. thaliana* NLR architectures. Expands to the architectures in fig. 3.2e those contributed by only one gene in the NLR'ome. Domain architectures are shown in the y-axis. The number of NLRs in each architecture is shown in x-axis. Asterisks indicate the 49 architectures not yet detected in the Brassicaceae family outside of *A. thaliana*, or in the reference accession Col-0. Numbers next to y-axis show the ratio of paired NLRs divided by the total number of NLRs in each architecture.

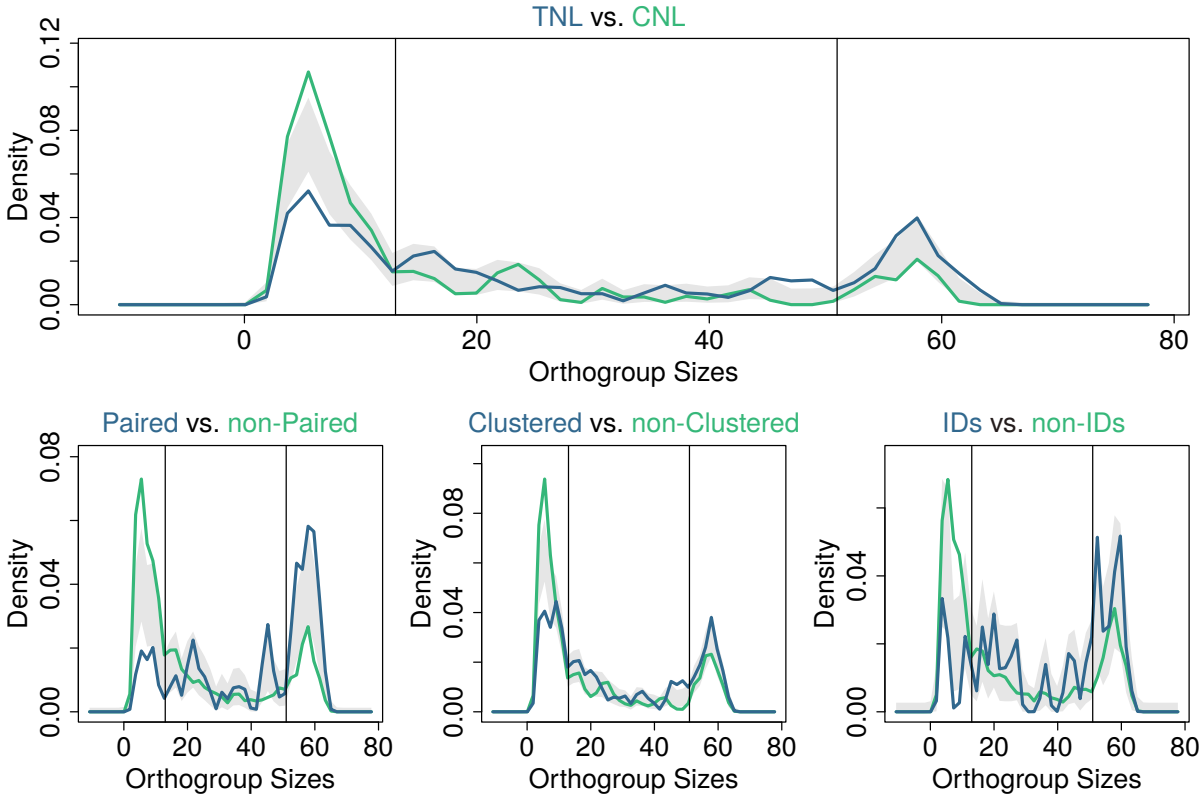


Figure 3.A.5.: OG size distribution comparisons
Vertical black lines divide cloud (left section) from shell (middle section) and core (right section) NLRs. a) Comparison of OG size distributions of TNL OGs (blue) and CNL OGs (green) b) Comparison of paired (blue) and non-paired (green) OGs c) Comparison of clustered (blue) and non-clustered (green) OGs d) Comparison of ID-containing (blue) and non-ID-containing OGs (green)

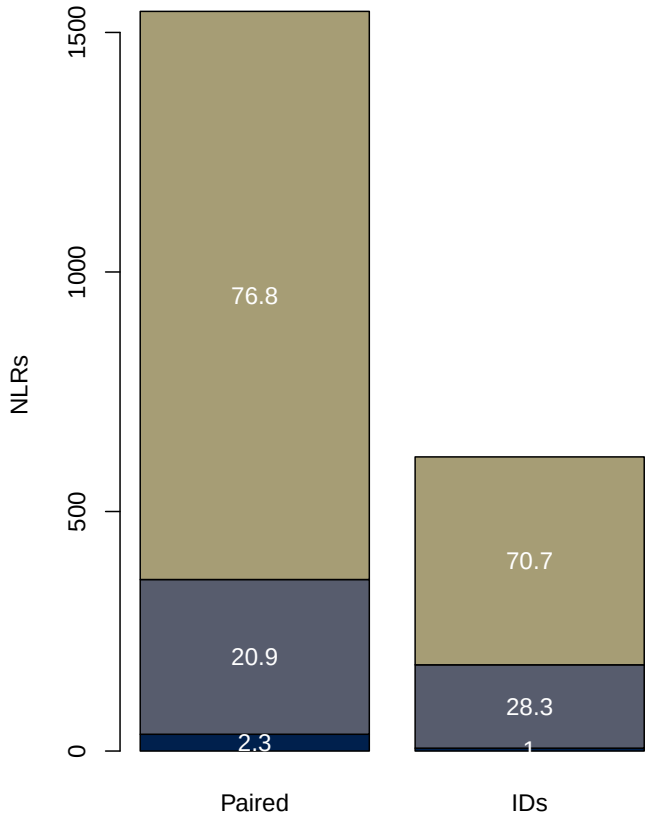


Figure 3.A.6.: Distribution of Paired NLRs and NLRs with IDs
Shown is the total number of NLRs in the Cloud (dark blue), the Shell (grey), and the Core (olive green), and the percentage (white text in the bars).

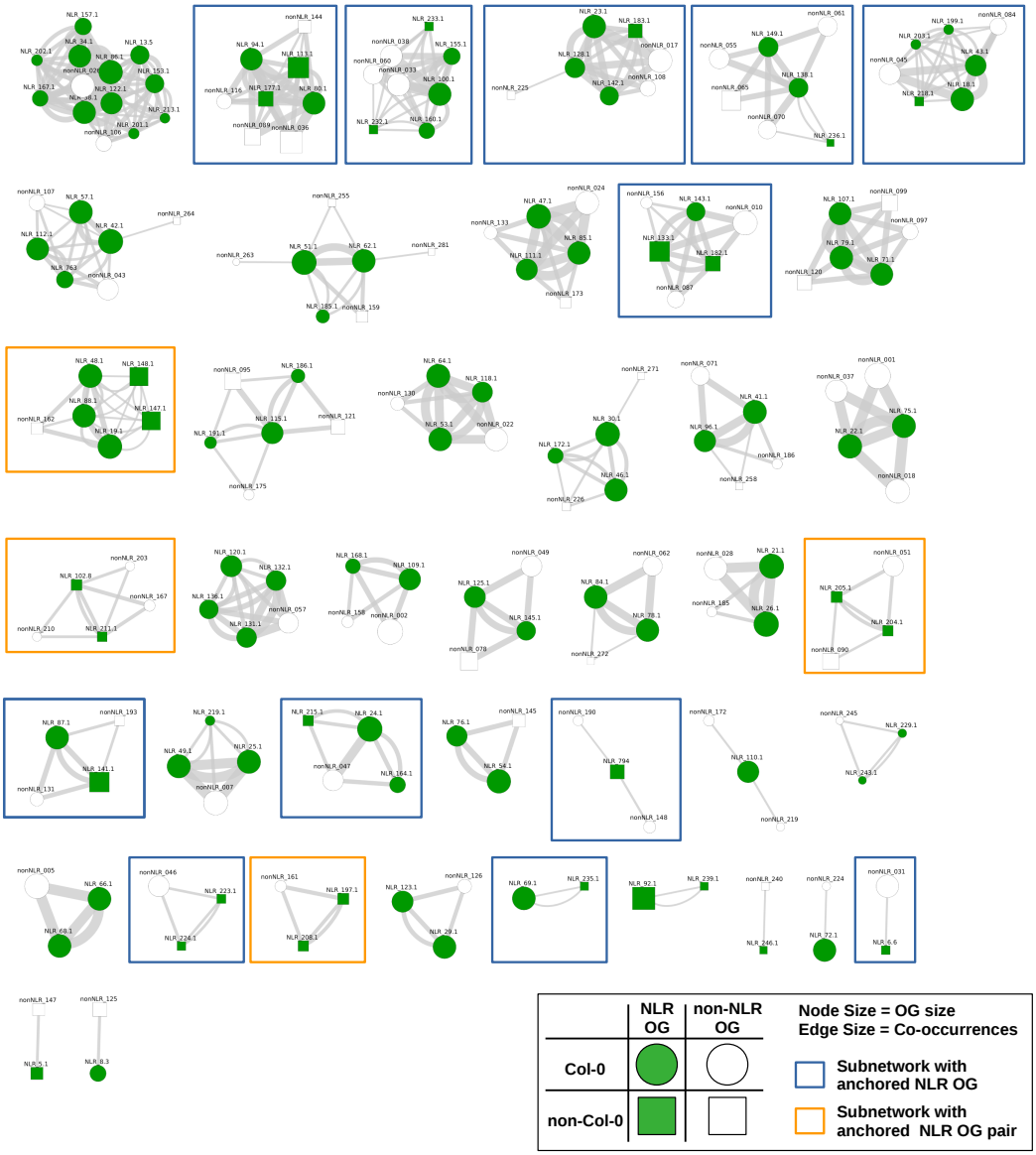


Figure 3.A.7.: Orthogroup (OG) co-occurrence network
Threshold set to 10 or more accessions. Annotated NLR (green nodes) and non-NLR genes (white nodes) clustered into OGs were analyzed for co-occurrence in the same contig. The number of co-occurrences is represented by grey lines connecting nodes (edges). The minimal co-occurrence threshold imposed was 10 accessions, but similar networks can be derived for any number accessions. NLR OGs without a Col-0 allele (green square nodes) are highlighted in blue boxes. Hypothetically paired OGs not known in Col-0 are highlighted in orange boxes.

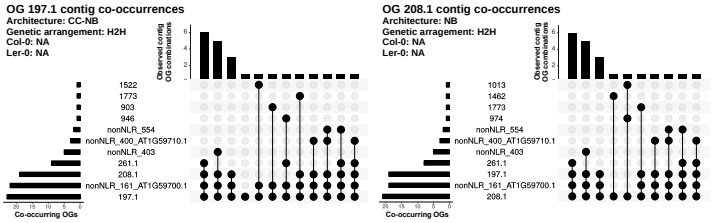


Figure 3.A.8.: Co-occurrence of OG197.1 and OG208.1
Quantitative co-occurrence of the novel hypothetical paired NLRs in OG197.1 and OG208.1. Abbreviations: OG, Orthogroup; H2H, Head-to-head; NA, Not available

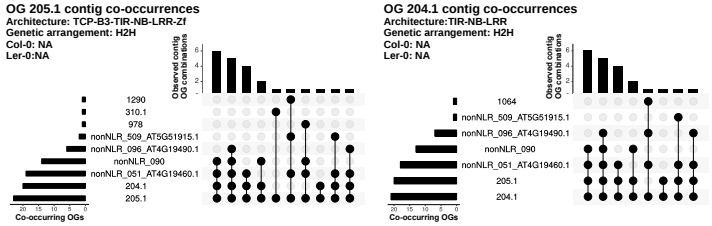


Figure 3.A.9.: Co-occurrence of OG205.1 and OG204.1
Quantitative co-occurrence of the novel hypothetical paired NLRs in OG205.1 and OG204.1. Abbreviations: OG, Orthogroup; H2H, Head-to-head; NA, Not available

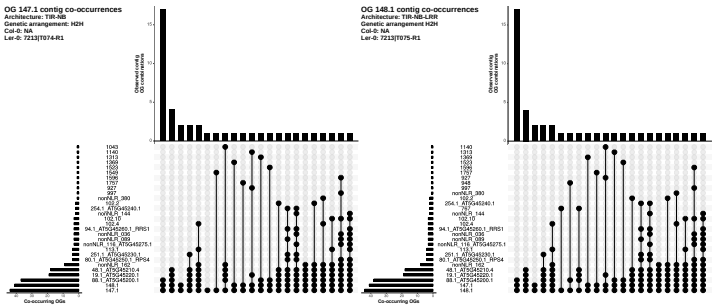


Figure 3.A.10.: Co-occurrence of OG147.1 and OG148.1
Quantitative co-occurrence of the novel hypothetical paired NLRs in OG147.1 and OG148.1. Abbreviations: OG, Orthogroup; H2H, Head-to-head; NA, Not available

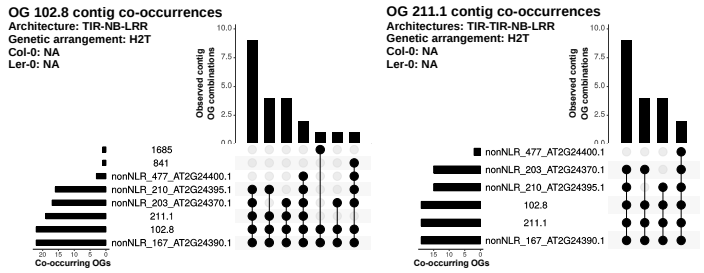


Figure 3.A.11.: Co-occurrence of OG102.8 and OG211.1
Quantitative co-occurrence of the novel hypothetical paired NLRs in OG102.8 and OG211.1. Abbreviations: OG, Orthogroup; H2T, Head-to-tail; NA, Not available

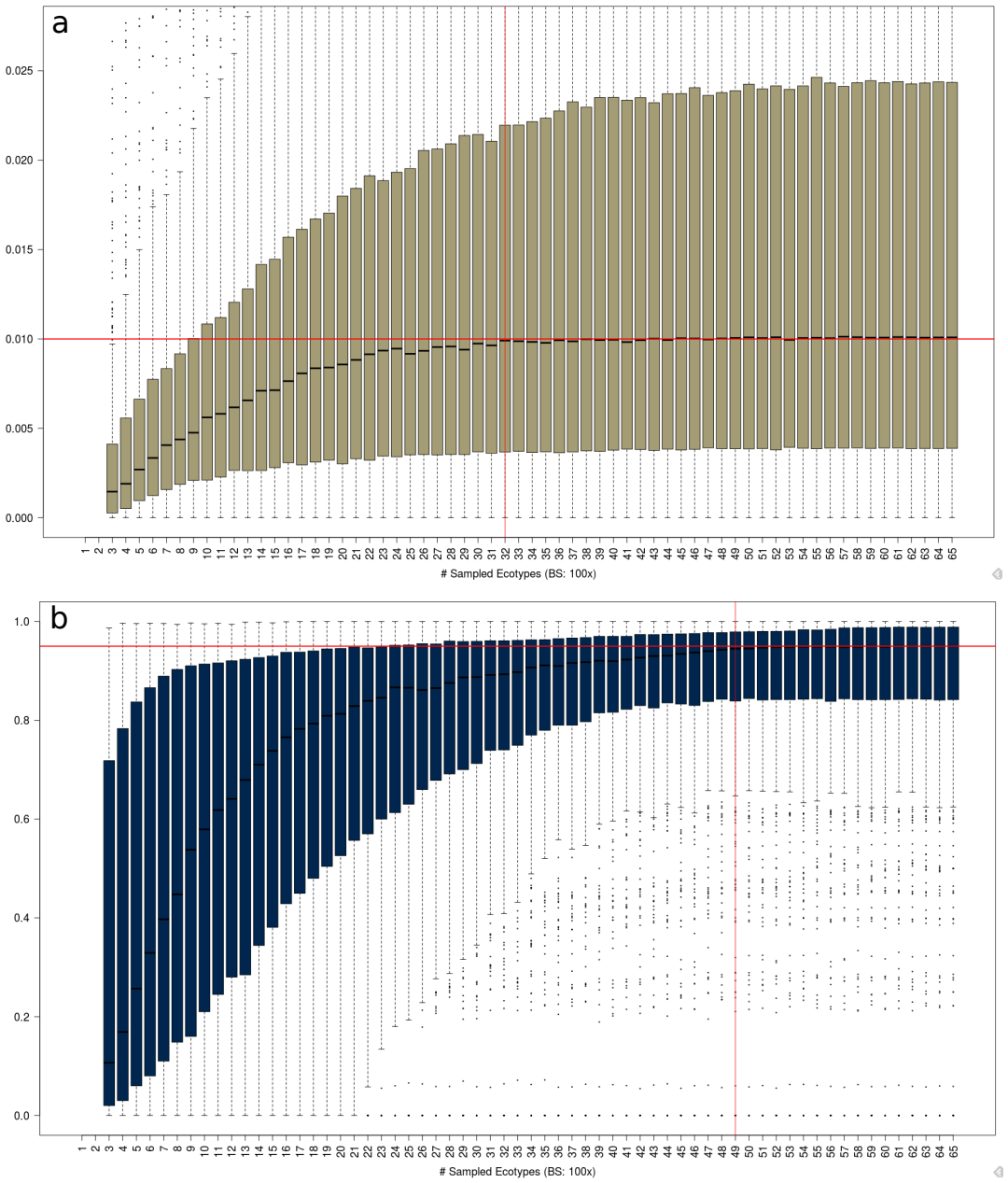


Figure 3.A.12.: Nucleotide and Haplotype Saturation
The red cross shows the saturation point. a) Nucleotide saturation. The y-axis shows the absolute value. b) Haplotype Saturation. The y-axis shows the saturation proportion.

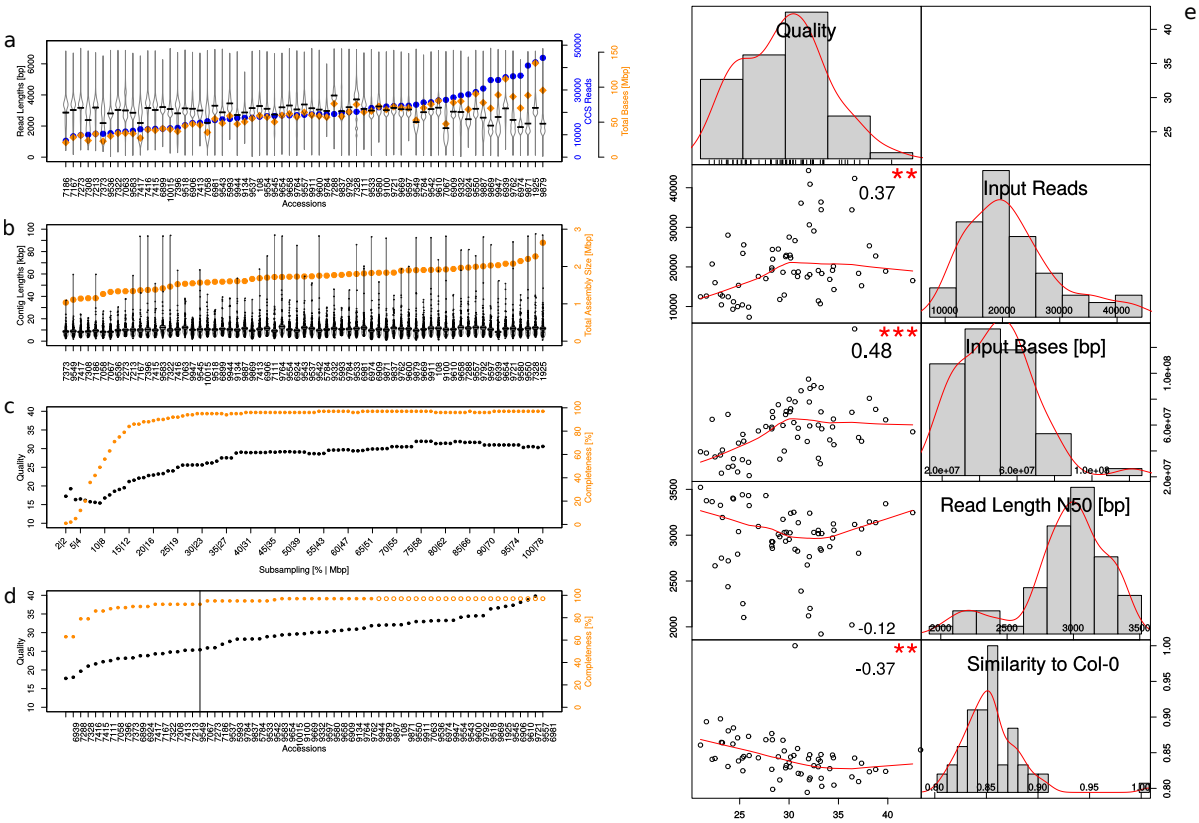


Figure 3.A.13.: Read and Assembly statistics
a) Read length distribution (Q20-filtered CCS reads) for all accessions (black). The mean is shown as a solid black horizontal line. The full densities are shown as a bean plot. The total number of CCS reads (blue circles) and the total number of bases (orange diamonds) are plotted in addition. b) Contig length distribution (black). The mean is shown as a solid black horizontal line and the 95 % Highest density Intervals (HDI: points in the interval have a higher probability than points outside) are shown as solid bands around the sample mean. The full densities are shown as a bean plot. Raw data points are plotted using black dots. The total assembly size (orange circles) is plotted in addition. c) Quality (black) and Completeness values (orange) for sub-sampled Col-0 datasets. The amount of input data for each sub-sampling experiment is shown as a second x axis. d) Quality (black) and Completeness values (orange) for all RenSeq accessions. Unfilled circles indicate accessions with qualities larger than any sub-sampled dataset. The vertical black line is drawn at 95 % Completeness. e) Correlations between the Assembly Quality, the amount of Input Reads, the amount of Input Bases [bp], the Read Length N50 [bp], and the Similarity to Col-0 are shown for the RenSeq datasets. Histograms and kernel densities (red line) are plotted for each variable. Scatter plots for variable pairs are shown together with a fitted line (red) and the Pearson's correlation coefficient (significance 0 '***',0.001 '**', 0.01 '*', 0.05 '.', 1 ' ').

3.B. Supplemental Tables

Table 3.B.1.: Used oligo sequences

oligo name	oligo sequence
NEBNext (E6861A)	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
NEBNext (E7350/E7335)	CAAGCAGAAGACGGCATACGAGATnnnnnnGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
AJI_1	CAAGCAGAAGACGGCATACGAGATcgttggttGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_2	CAAGCAGAAGACGGCATACGAGATttctggttGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_3	CAAGCAGAAGACGGCATACGAGATtggcggttGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_4	CAAGCAGAAGACGGCATACGAGATtagtcgttGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_5	CAAGCAGAAGACGGCATACGAGATtggttcttGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_6	CAAGCAGAAGACGGCATACGAGATagagttctGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_7	CAAGCAGAAGACGGCATACGAGATtccattggGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_8	CAAGCAGAAGACGGCATACGAGATccagctggGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
AJI_9	CAAGCAGAAGACGGCATACGAGATgcagacggGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
OF-PacF1	aatgatacggcgaccaccgaGATcgttggttCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF2	aatgatacggcgaccaccgaGATttctggttCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF3	aatgatacggcgaccaccgaGATtggcggttCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF4	aatgatacggcgaccaccgaGATtagtcgttCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF5	aatgatacggcgaccaccgaGATtggttcttCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF6	aatgatacggcgaccaccgaGATagagttctCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF7	aatgatacggcgaccaccgaGATtccattggCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF8	aatgatacggcgaccaccgaGATccagctggCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
OF-PacF9	aatgatacggcgaccaccgaGATgcagacggCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
Illumina_P5	AATGATACGGCGACCACCGA
Illumina_P7	CAAGCAGAAGACGGCATACGAGAT

Table 3.B.2.: The table shows for each accession the used 'Identifier' and the corresponding 1001 Identifier (1001 id). The accession name (Accession) is given, as well as the seed stock numbers if known (Stock). The country (Origin) of each accession is given together with the coordinates (lat and long). It is shown if an accession belongs to the 'Relict' group or is a founder of the MAGIC-lines (MAGIC).

Identifier	1001_id	Accession	Stock	Origin	lat	long	Relict	MAGIC
5784	5784	Ty-1	CS78790	UK	56.4	-5.2	0	0
6981	6981	Ws-2	CS76631	RUS	52.3	30	0	0
9134	9134	Yeg-8	CS75475	ARM	39.87	45.36	0	0
9610	9610	Lesno-4	CS77034	RUS	53.04	51.96	0	0
10015	10015	Sha	CS22690	AFG	37.29	71.3	0	0
5993	5993	DraIV 6-22	CS76823	CZE	49.41	16.28	0	0
9669	9669	Mitterberg-2-185	CS77086	ITA	46.37	11.28	0	0
9784	9784	Erg2-6	CS76845	GER	48.5	8.8	0	0
9792	9792	Lu4-2	CS77058	GER	48.54	9.09	0	0
1925	1925	MNF-Che-2	CS76185	USA	43.53	-86.18	0	0
6909	6909	Col-0	CS22681	USA	38.3	-92.3	0	1
9100	9100	Lag1-2	CS75441	GEO	41.83	46.28	0	0
9658	9655	Marce-1	#N/A	ITA	38.92	16.47	0	0
9721	9721	Schip-1	CS77239	BUL	42.72	25.33	0	0
9533	9533	Cem-0	CS76763	ESP	41.15	-4.32	1	0
9542	9542	Fun-0	CS76872	ESP	40.79	-4.05	1	0
9550	9550	Iso-4	CS7694	ESP	43.05	-5.37	1	0
9554	9554	Lso-0	CS77055	ESP	38.86	-3.16	1	0
9600	9600	Vis-0	CS78848	ESP	39.85	-6.04	1	0
9518	9518	Alm-0	CS76660	ESP	39.88	-0.36	1	0
9537	9537	Cum-1	CS76787	ESP	38.07	-6.66	0	0
9557	9557	Moa-0	CS77102	ESP	42.46	0.7	0	0
9597	9597	Vig-1	CS78843	ESP	42.31	-2.53	0	0

Table 3.B.2.: continued

Identifier	1001_id	Accession	Stock	Origin	lat	long	Relict	MAGIC
6899	6899	Bay-0	CS22676	GER	49	11	0	0
6906	6906	C24	CS22680	POR	40.21	-8.43	0	0
6911	6911	Cvi-0	CS76789	CPV	15.11	-23.62	1	0
9580	9580	Scm-0	CS77241	ESP	38.68	-3.57	0	0
9654	9654	Liri-1	CS77041	ITA	41.41	13.77	0	0
108	108	LDV-18	CS77013	FRA	48.52	-4.07	0	0
9911	9928	BEZ-9	#N/A	FRA	44.12	3.77	0	0
6981.2	6981	Ws-2	CS28828	RUS	52.3	30	0	0
7058	7058	Bur-0	CS28124	IRL	54.1	-6.2	0	1
7111	7111	Edi-0	CS28220	UK	55.95	-3.16	0	1
7213	7213	Ler-0	CS28445	GER	47.98	10.87	0	1
7288	7288	Oy-0	CS28591	NOR	60.39	6.19	0	1
7373	7373	Tsu-0	CS28780	JPN	34.43	136.31	0	1
7067	7067	Ct-1	CS28195	ITA	37.3	15	0	1
7186	7186	Kn-0	CS28395	LTU	54.9	23.89	0	1
7273	7273	No-0	CS28565	GER	51.06	13.3	0	1
7396	7396	Ws-0	CS28824	RUS	52.3	30	0	1
7413	7413	Wil-2	TSL-JJ-SP2486	LTU	54.68	25.32	0	1
6909.2	6909	Col-0	CS28167	USA	38.3	-92.3	0	0
7322	7322	Rsch-4	CS28716	RUS	56.3	34	0	1
7415	7415	Wu-0	N6897	GER	49.79	9.94	0	1
7416	7416	Yo-0	CS22624	USA	37.45	-119.35	0	0
7063.2	7063	Can-0	CS28130	ESP	29.21	-13.48	1	1
7328	7328	Sf-2	CS28731	ESP	41.78	3.03	0	1
6939	6939	Mt-0	N1380	LIB	32.34	22.46	0	1
7167	7167	Hi-0	CS28346	NED	52	5	0	1
7308	7308	Po-0	CS28648	GER	50.72	7.1	0	1
6924	6924	HR-5	CS22596	UK	51.41	-0.64	0	0
7417	7417	Zu-0	N6902	SUI	47.37	8.55	0	1
9536	9536	Cor-0	CS76782	ESP	40.83	-2	1	0
7063	7186	Kn-0	#N/A	LTU	54.9	23.89	0	1
9543	9543	Gra-0	CS76886	ESP	36.77	-5.39	1	0
9545	9545	Her-12	CS76920	ESP	39.4	-5.78	1	0
9549	9549	Hum-2	CS76943	ESP	42.23	-3.69	1	0
9583	9583	Sne-0	CS77258	ESP	37.09	-3.38	1	0
9837	9837	Con-0	CS76780	ESP	37.94	-5.6	1	0
9871	9871	Nac-0	CS77117	ESP	40.75	-3.99	1	0
9944	9944	Don-0	CS76411	ESP	36.83	-6.36	1	0
6974	6974	Ull2-5	CS78818	SWE	56.06	13.97	0	0
9555	9555	Mar-1	CS77068	ESP	39.58	-3.93	1	0
9598	9598	Vim-0	CS78844	ESP	41.88	-6.51	1	0
9905	9905	Ven-0	CS78840	ESP	40.76	-4.01	1	0
9762	9762	Etna-2	CS76487	ITA	37.69	14.98	1	0
9764	9764	Qar-8a	CS76581	LBN	34.1	35.84	1	0
9332	9332	Bar-1	CS76688	SWE	62.87	18.38	0	0
9869	9869	Moj-0	CS77105	ESP	36.76	-5.28	1	0
9879	9879	Per-0	CS77169	ESP	37.6	-1.12	1	0
9887	9887	Pun-0	CS77196	ESP	40.4	-4.77	1	0
9947	9947	Ped-0	CS76415	ESP	40.74	-3.9	1	0
9832	9832	Cat-0	CS76759	ESP	40.54	-3.69	1	0

Table 3.B.3.: The table shows for each accession (Identifier),which size selection method was used (Size_Sel: BP=BluePippin, SE=SageElf). The Sequencing Provider (Seq_Prov) (MPI=Max Planck Institute for Developmental Biology, Tuebingen, EI=Earlham Institute Norwich, UNC=University of Chapel Hill) and the Sequencing Facility (Seq_Fac) are given, and the used 'Library Adaptors' are shown (for custom adaptor sequences see table 3.B.1). The table contains the number of sequenced SMRT cells (cells), and if an accession was sequenced multiplexed (multi). It also shows for which accessions PCRfree whole genome sequencing short read data (SR) was produced.

Identifier	Size_Sel	Seq_Prov	Seq_Fac	Library_Adaptors	cells	multi	SR
5784	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6981	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9134	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9610	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7351	1	No	1
10015	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
5993	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9669	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7354	1	No	1
9784	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9792	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
1925	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	2	No	1
6909	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9100	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9658	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7353	1	No	1
9721	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7355	1	No	1
9533	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9542	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9550	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9554	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9600	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9518	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9537	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9557	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9597	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6899	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6906	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6911	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	NA
9580	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9654	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7352	1	No	1
108	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	2	No	1
9911	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6981.2	SE	EI	TSL	AJI_6+OF-PacF6		Yes	NA
7058	SE	EI	TSL	AJI_1+OF-PacF1	8	Yes	1
7111	SE	EI	TSL	AJI_8+OF-PacF8	5	Yes	1
7213	SE	EI	TSL	AJI_3+OF-PacF3	5	Yes	1
7288	SE	EI	TSL	AJI_1	1	No	1
7373	SE	EI	TSL	AJI_5+OF-PacF5	8	Yes	1
7067	SE	EI	TSL	AJI_3+OF-PacF3	8	Yes	1
7186	SE	EI	TSL	AJI_2+OF-PacF2	5	Yes	1
7273	SE	EI	TSL	AJI_4+OF-PacF4	5	Yes	1

Table 3.B.3.: continued

Identifier	Size_Sel	Seq_Prov	Seq_Fac	Library_Adaptors	cells	multi	SR
7396	SE	EI	TSL	AJI_6+OF-PacF6	8	Yes	1
7413	SE	EI	TSL	AJI_1	1	No	1
6909.2	SE	EI	TSL	Nextera1F+1R			NA
7322	SE	EI	TSL	AJI_5+OF-PacF5	5	Yes	1
7415	SE	EI	TSL	AJI_7+OF-PacF7	5	Yes	1
7416	SE	EI	TSL	AJI_1			1
7063.2	SE	EI	TSL	AJI_2+OF-PacF2	1	Yes	NA
7328	SE	EI	TSL	AJI_1	1	No	1
6939	SE	EI	TSL	AJI_9+OF-PacF9	5	Yes	1
7167	SE	EI	TSL	AJI_1+OF-PacF1	5	Yes	1
7308	SE	EI	TSL	AJI_4+OF-PacF4	8	Yes	1
6924	SE	EI	TSL	AJI_1	1	No	1
7417	SE	EI	TSL	AJI_7+OF-PacF7	8	Yes	1
9536	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
7063	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9543	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9545	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9549	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9583	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9837	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9871	SE	UNC	UNC-HTSF	NEBNext_E7335	2	No	1
9944	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
6974	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9555	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	NA
9598	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	NA
9905	SE	UNC	UNC-HTSF	NEBNext_E7335	2	No	NA
9762	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9764	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9332	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9869	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	4	No	1
9879	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9887	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9947	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9832	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	4	No	NA

Table 3.B.4.: Misannotated NLRs

ID	Reason for Exclusion
7328/G707	misannotation
9554/G142	misannotation
9658/G071	misannotation
9669/G017	misannotation
7396/G336	misannotation
9332/G195	misannotation
7273/G160	misannotation
6924/G055	misannotation
9654/G214	misannotation

Table 3.B.5.: Tajima's D comparison for NLR pairs.

araport11	synonym	pairclass	idnew	Tajima.D	araport11	synonym	pairclass
AT5G17880.1	*CSA1/WRR5a*	executor	130.1	2.313171443	*AT5G17890.1*	*CHS3/DAR4*	sensor
AT5G45060.1	*RPS4B*	executor	43.1	-0.69202534	*AT5G45050.1*	*RRS1B*	sensor
AT5G45250.1	*RPS4*	executor	94.1	-0.902908159	*AT5G45260.1*	*RRS1*	sensor
NA	NA	executor	148.1	-2.475406708	NA	NA	sensor
NA	NA	executor	205.1	0.329520807	NA	NA	sensor
AT1G72840.2	NA	executor	13.5	3.246254466	*AT1G72850.1*	NA	sensor
AT1G72860.2	NA	executor	213.1	-1.073582225	*AT1G72870.1*	NA	sensor
AT5G48770.1	NA	executor	243.1	-0.797767387	*AT5G48780.1*	NA	sensor
AT5G45230.1	NA	executor	254.1	-1.318633405	*AT5G45240.1*	NA	sensor
AT4G36150.1	NA	executor	21.1	-0.093731171	*AT4G36140.1*	NA	sensor
AT3G51570.1	NA	executor	84.1	-2.416786807	*AT3G51560.1*	NA	sensor
AT5G40100.1	NA	executor	67.1	-2.536090992	*AT5G40090.1*	NA	sensor
AT2G17060.1	NA	executor	106.1	-2.653985925	*AT2G17050.2*	NA	sensor
AT1G17600.1	*SOC3*	executor	47.1	-2.563305714	*AT1G17610.1*	*CHS1*	sensor
AT5G45200.1	NA	executor	48.1	-1.127247783	*AT5G45210.4*	NA	sensor
AT4G12010.1	*DSC1*	executor	41.1	0.169959165	*AT4G12020.2*	*MAPKKK11*	sensor
AT1G12210.1	*RFL1*	control	168.1	-2.49472187	*AT1G12220.1*	*RPS5*	control
AT1G12280.1	*summ2*	control	68.1	-2.499388809	*AT1G12290.1*	NA	control
AT1G27180.1	NA	control	24.1	-2.111599699	*AT1G27170.1*	NA	control
AT1G56510.1	*WRR4*	control	140.1	-2.180811584	*AT1G56520.2*	NA	control
AT1G63860.1	NA	control	142.1	0.337557903	*AT1G63870.1*	NA	control
AT1G72900.1	NA	control	34.1	-1.264307713	*AT1G72890.2*	NA	control
AT1G72910.1	NA	control	202.1	-1.881193865	*AT1G72920.1*	NA	control
AT1G72930.1	NA	control	86.1	0.415710515	*AT1G72940.1*	NA	control
AT4G19050.1	NA	control	76.1	-0.513451551	*AT4G19060.1*	NA	control
AT5G18350.1	NA	control	46.1	-1.743490776	*AT5G18360.1*	NA	control
AT5G38340.1	NA	control	53.1	0.323140281	*AT5G38350.1*	NA	control
AT5G41540.1	NA	control	138.1	-1.715007139	*AT5G41550.1*	NA	control
AT5G44900.1	NA	control	131.1	-1.890234716	*AT5G44910.1*	NA	control
AT5G45070.1	NA	control	203.1	-0.967246358	*AT5G45080.1*	NA	control
AT5G47250.1	NA	control	49.1	1.935455359	*AT5G47260.1*	NA	control

Table 3.B.6.: CC detection in known functional CNLs
CC predictions (binary, 0/1) from `Paircoil2`, `Coils`, and the `NLR-parser` for known functional CNLs. The `Araport11` identifier and the name of each CNL is given.

Identifier	Name	`Paircoil2`	`Coils`	`NLR-parser`
AT3G07040	*RPM1/RPS3*	0	0	1
AT4G26090	*RPS2*	1	0	1
AT1G10920	*LOV1*	0	0	1
AT1G58602	*RPP7*	1	0	1
AT1G12220	*RPS5*	0	1	1
AT5G43470	*HRT/RPP8*	1	1	1
AT1G33560	*ADR1*	0	0	0
AT1G12280	*summ2*	1	1	1
AT1G12210	*RFL1*	1	1	1
AT1G59620	*CW9*	0	0	1
AT1G61180	*Uni-1d (Ws)*	1	1	1
AT1G61190	*RPP39*	1	1	1
AT3G46530	*RPP13*	0	1	1
AT3G50950	*ZAR1*	1	1	1

Table 3.B.7.: Brassicaceae species used for domain comparisons.

year	version	Species	NCBI_Species	Subspecies	NCBI_Subspecies
2016	2016-11-10	*Leavenworthia alabamica*	310722		
2016	2016-11-10	*Aethionema arabicum*	228871		
2016	2016-11-10	*Schrenkiella parvula*	98039		
2017	GCA_000733195.1	*Arabis alpina*	50452		
2016	2016-11-10	*Sisymbrium irio*	3730		
2016	173	*Eutrema halophilum*	98038		
2016	29	*Eutrema salsugineum*	72664		
2017	1	*Raphanus sativus*	3726	*Raphanus sativus var. hortensis*	
2016	1.5	*Brassica rapa*	3711	*Brassica rapa subsp. pekinensis*	51351
2017	1.1	*Brassica nigra*	3710		
2016	5	*Brassica napus*	3708		
2017	1.1	*Brassica juncea*	3707		
2017	GCF_000695525.1	*Brassica oleracea*	3712	*Brassica oleracea var. oleracea*	109376
2016	1.1	*Brassica oleracea*	3712	*Brassica oleracea var. capitata*	3716
2016	2	*Camelina sativa*	90675		
2016	1.1	*Capsella grandiflora*	264402		
2016	1	*Capsella rubella*	81985		
2017	GCA_900078215.1	*Arabidopsis halleri*	81970	*Arabidopsis halleri subsp. gemmifera*	63677
2017	BASO01	*Arabidopsis halleri*	81970	*Arabidopsis halleri subsp. gemmifera*	63677
2016	1	*Arabidopsis lyrata*	59689	*Arabidopsis lyrata subsp. lyrata*	81972
2017	BASP01	*Arabidopsis lyrata*	59689	*Arabidopsis lyrata subsp. petraea*	59691
2017	carhr38	*Cardamine hirsuta*	50463		

3.C. Supplemental Material

3.C.1. Re-annotation SOP

SOP Re-Annotation

http://ann-nblrrome.tuebingen.mpg.de

Warnings

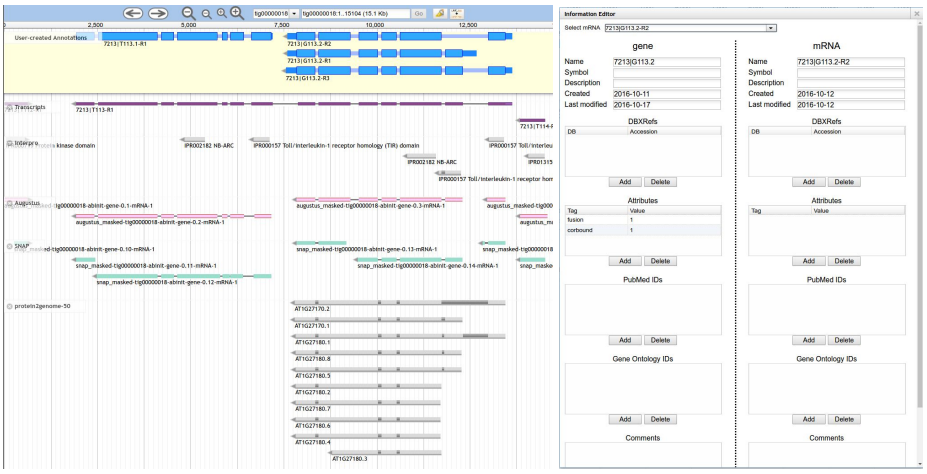
WebApollo does allow duplicate names!

Comments

We only re-annotate genes with NB- or TIR- domains

Normal Re-annotation

1. Check if regions with evidence (prefer ESTs > Proteins > Gene Predictions) have a gene model (if not add it)
2. Check the gene model (exon/intron structure) against protein and EST evidence (always compare SNAP and Augustus in case of likely fusions)
3. Re-annotate if necessary
    a. Add UTR information from est2genome track if it can be easily incorporated into the gene model
    b. Compare exon-intron structures from transcript and protein evidence to maker annotation and adjust where needed
4. Rename to original transcript annotation of MAKER (example: gene name=6909|G040, transcript name=6909|T040-R1)
5. If several transcripts are annotated, use -R1, -R2, etc
   Example: Gene 7213|G113
   Additionally to being split (see ‘Gene Fusion Handling’), three transcripts are annotated: 7213|T113.2-R1, 7213|T113.2-R2 and 7213|T113.2-R3



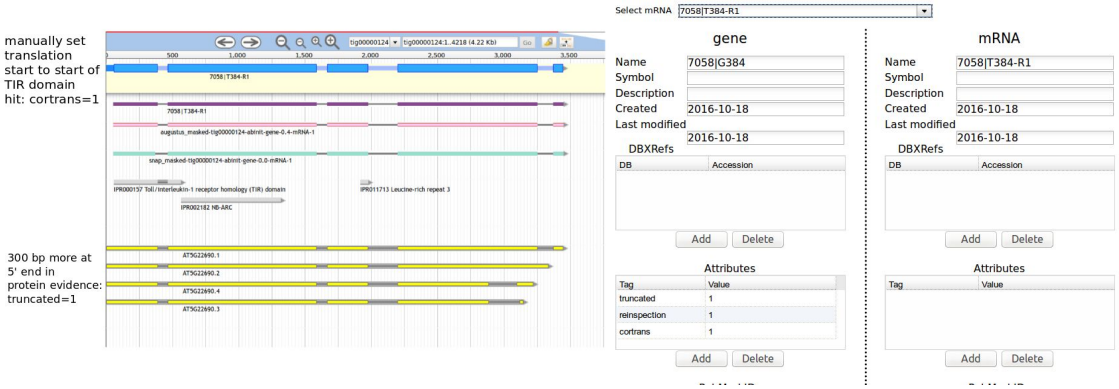
6. Set appropriate tag-value pairs (see ‘Attribute Settings’)

Truncated Genes

1. Check NLR genes at the beginning or end of contigs (especially if the distance to the contig border is < 500bp)
2. Check if available evidence is longer than annotated (causes alignment overhangs)
3. If possible, correct the annotation using the evidence
    a. If you correct exon-/intron- boundaries, set attribute corbound=1
    b. If you correct translation start or end, set attribute cortrans=1

    (!) Be careful with those adjustments and do only when evidence is bulletproof.
4. Add attribute truncated=1

Example: Gene 7058|G384



Splitting Gene Fusions

1. Split gene according to evidence
2. Rename the new genes:
   General Format: ACC|[G/T]*XXX*.*Y*-RZ
    - ACC=Accession ID
    - [G/T]*XXX*=Gene/Transcript ID
    - .*Y*=Split Gene/Transcript Sub-ID
    - -RZ=Isoform ID

   Example: Gene 6909|G040
    - Original gene name 6909|G040 → 6909|G040.1 and 6909|G040.2 (the leftmost gene gets appendix '.1', then '.2', '.3' etc follow while moving right)
    - Original transcript name 6909|T040-R1 → 6909|T040.1-R1 and 6909|T040.2-R1
3. Add attribute fusion=1

Genes Without Evidence

- This situation is common for non-reference datasets. You might observe many predictions without EST or Protein evidences. Besides using BLAST, the Araport JBrowser (https://apps.araport.org/jbrowse/?data=arabidopsis) is handy to verify/compare gene architectures. In some cases these instances are pseudogenes, and so the Araport information is useful. Sequences, or gene identifiers can be searched directly by using the field left to the Go button in the top of the browser.

- If you find a pseudogene (based on Araport annotation and synteny), you can flag it with pseudogene=ATG_identifier.
- The noevidence=1 parameter is only set if we choose to reannotate something without evidence (evidence here is everything that gives you a hint for this reannotation to be necessary, so no evidence at all will probably nearly never be the case)

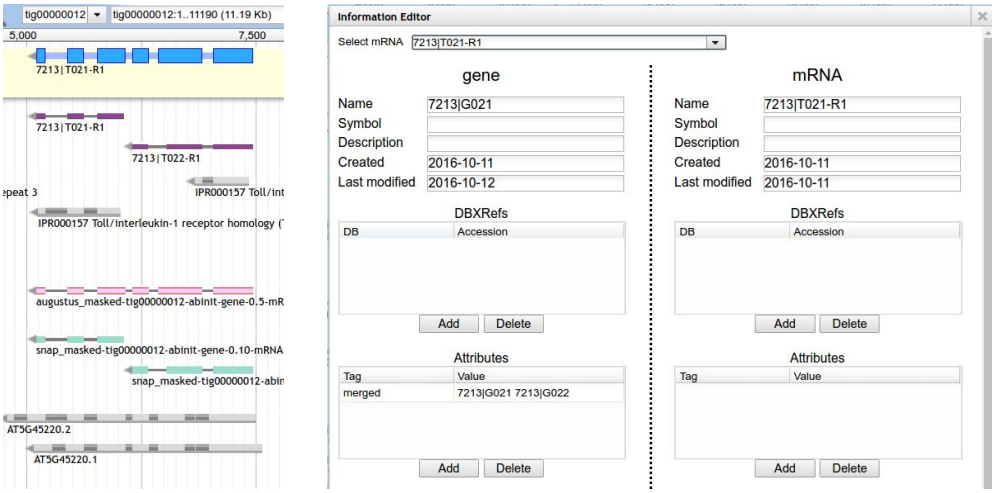
Merging Split Genes

Be very careful when merging genes. It is worse for orthogroup predictions to have a wrongly merged gene, than to have two erroneously split genes.

1. Merge
2. Rename the merged gene according to the 'leftmost' gene that is contained
3. Add the tag 'merged' with all the gene names that are merged. Separate by space (do not use comma)

Example: Genes 7213|G021 and 7213|G022

- New gene name: 7213|G021
- New transcript name: 7213|T021-R1
- add tag: merged='7213|G021 7213|G022' (Do not use comma as separator)



Example: Merging a gene with another one that had been flagged as part of a fusion

- Original gene names: 7058|G069 and 7058|G070
- After splitting: 7058|G069.1 (fusion=1), 7058|G069.2 (fusion=1), and 7058|G070
- After merging: 7058|G069.1 (fusion=1) and 7058|G069.2 (fusion=1; merged=7058|G069.2 7058|G070)

Annotating New Genes

In case we want to add an additional new annotation.

1. Name the new gene according to the gene to its left, and add ".N<number>"
   Example: New gene next to 1925|G530
   - Gene Name 1925|G530.N1
   - mRNA Name 1925|T530.N1-R1

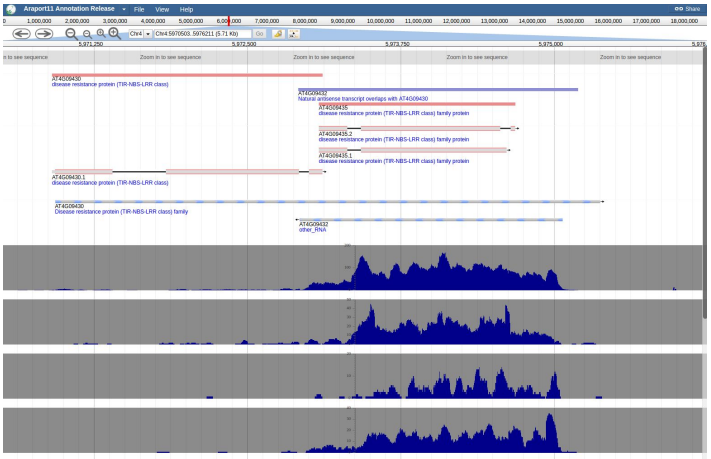
Attribute Settings

Tag	Value	
reinspection	1	set if reinspection of the gene model is needed
fusion	1	set for split genes. Each gets the fusion tag.
truncated	1	set if a gene seems to be truncated
pair	GeneID of “partner”	set if evidence is in “Col-0 Pairs” from pairs_and_putpairs list
putpair	gene ID of “partner” gene	set if head-to-head orientation is given and evidence is not in the “Col-0 Pairs” from pairs_and_putpairs list
pseudogene	Araport11 Identifier	set if there is pseudogene evidence from Col-0
noevidence	1	set if the reannotation was done without any source of evidence (nearly never the case)
merged	GeneID<space>GeneID(<space>GeneID...)	set for merged gene
corbound	1	set if exon-/intron- boundaries were changed without direct evidence
cortrans	1	set if translation start or end was set without direct evidence
misassembly	1	set if a misassembled contig is suspected
delete	1	use if a gene model should not be replaced, but deleted completely
mod	1	use if gene model was extensively changed (mostly without evidence from gene predictors or transcript/protein mappings) in order to rescue the domain structure. Genes that were re-annotated this way probably need to be excluded from some analyses.

Further Remarks:

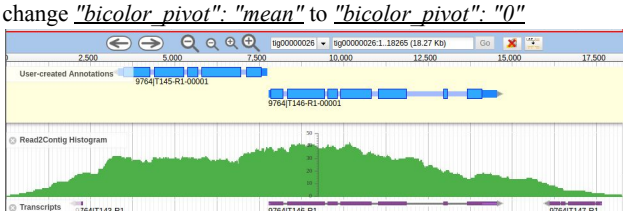
- We found a misannotation in Araport11. They split the tair10 gene AT4G09430 into AT4G09430 and AT4G09435, both annotated as TNLs. This is wrong. The first gene contains the NB- and the TIR- domain, the second one only has a LRR annotated. The misannotation is driven by a natural antisense gene AT4G09432 that overlaps AT4G09430. This antisense gene is expressed, whereas AT4G09430 might not be expressed (at least not in their data). Araport11 mistreats the expression as belonging to the TNL, and splits the gene. Long story short, don’t split a gene if your evidence is from AT4G09430 and/or AT4G09435 :)
  For a visual reference check: 6909|G370

- Be careful if you have to re-annotate two overlapping genes. They always get treated as two isoforms from the same gene. We agreed on removing UTR that is overlapping.



- Change default view of **Read2Contig Histogram** track.

change *"bicolor_pivot": "mean"* to *"bicolor_pivot": "0"*



- Noncanonical splice sites:
  found e.g. in AT5G47280, AT4G33300

4. Discussion and Outlook

My PhD project is embedded in a collaborative effort centered around using state-of-the-art targeted long read sequencing (SMRT RenSeq) to pin down the importance of NLRs in plant immunity (2Blades Foundation project 'Resistance Gene Diversity', 2Blades (2015)).

Precious RenSeq studies successfully improved the annotation of known NLR genes, and detected unreported NLRs in several species (see section 1.5.1). Enriching a sample for the genomic proportion that contains NLRs, facilitates the assembly by reducing the assembly complexity. Furthermore, focusing on sequencing only the NLR gene family reduces costs, which is especially important for research projects in species with bigger or complex genomes. RenSeq has mainly been used with short read Illumina sequencing to reevaluate existing NLR gene annotations, and to annotate previously unreported NLRs in the Solanaceae (Andolfo et al. 2014; Jupe et al. 2013). The method has further proven useful in an investigation of polymorphisms and evolutionary pressures in known NLRs of *S. pennellii* (Stam et al. 2016). Analyses focusing on the genetic mapping of a resistance trait of interest and on the completeness validation of known functional NLRs demonstrated RenSeq as a cost-efficient application for resistance breeding programs in crops (Armstrong et al. 2018; Chen et al. 2018; Jiang et al. 2018; Jupe et al. 2013; Van Weymers et al. 2016).

Here, RenSeq data of 65 *A. thaliana* accessions were used for the analysis of the species' pan-NLR'ome. Chapter 3 contains the unpublished manuscript The *Arabidopsis thaliana* pan-NLR'ome, which reports and analyzes the biological variability of NLRs in *A. thaliana*. With the help of colleagues, I found that known integrated domains (IDs) from the Brassicaceae are frequently shuffled in *A. thaliana* NLRs, which increases their architectural diversity, and additional novel IDs and architectures were defined. The pan-NLR'ome showed saturation, which allowed to define the core NLR genes, and to analyze presence/absence polymorphisms in non-core NLRs. Furthermore, I show haplotype saturation, quantify the selective forces that act on specific NLRs and domains, and detect evolutionary coupled co-evolving NLRs. A detailed discussion of the results is presented in the manuscript section 3.4 and section 3.5.

The results and the easy-to-adopt analysis pipeline can be used to support other RenSeq studies. Research in important crop species, like the ones that will be analyzed as part of the 2Blades project, will profit from the knowledge that was gained. In chapter 2, I describe the extensive method development and optimization of my PhD work, which was coupled tightly with sophisticated methods for quality control. Using a non-standard and state-of-the-art sequencing method like SMRT RenSeq required customized solutions for the contig assembly and the gene- and domain annotation. I showed that the combination of optimized automated methods for assembly and anno-

tation together with manual curation of NLR genes provides a reliable and complete representation of the NLR complements of 65 *A. thaliana* accessions.

The method optimization process and the best-practice results will be discussed in this chapter. I will examine the influence the RenSeq input datasets have on the assembly, the NLR annotation, and the analysis of NLR variation in *A. thaliana*. Furthermore, I discuss the benefit of WGS short and long read data for the assembly process, the gene annotation and for the quality control. Manual curation of automated gene annotations was a time consuming but necessary task that should be made superfluous in future research if possible. I will discuss possible improvements of the automated gene annotation, and also suggest ways to make manual curation less tedious.

The pan-NLR'ome data created during this PhD project was already used in other NLR-related research projects, as intended. The current and potential future use will be discussed in section 4.4.

4.1. RenSeq input critically influences the outcome of an NLR'ome project

The disadvantages of using RenSeq with short Illumina reads have been displayed in the introduction (section 1.4.2). Mainly the assembly of NLRs and NLR clusters suffers when using short read data. High fragmentation and incorrect NLR cluster assembly was reported (Andolfo et al. 2014; Jupe et al. 2013). These problems were solved in this study by using RenSeq with the long read sequencing SMRT technology from PacBio. The sequenced fragments were ∼3kb long (most between 2-5kb, fig. 2.8). This was even longer than the 2.5 kb-3.5 kb long fragments that have already been successfully used to assemble full length NLRs, clusters, and surrounding regions in potato (Witek et al. 2016b).

Witek et al. (2016b) include a short analysis of the influence that input data has on assembled NLRs. They sequenced a susceptible and a resistant *S. americanum* accession on three SMRT cells each, aiming at an NLR coverage between 30-50x. The susceptible accession contained fragments with lengths from 1.5 kb to 2.5 kb, whereas the resistant accession contained fragments between 3-4 kb length. They created reads of insert (ROIs) with minimum three passes and a maximum error of 10 %. The ROI construction is comparable to the CCS read construction procedure of my PhD study, but allowed more errors (max 1 % error was allowed in the pan-NLR'ome project). *De novo* assemblies were created using `Geneious`. A side note: this assembler did not perform well in initial assembly tests done during my PhD study. A Col-0 test showed for example many artificially duplicated genomic regions. I did not report in-depth results, because extensive testing of different assemblers was out of the scope of the PhD project.

As expected, assembled contig sizes increased with increasing input fragment lengths. Also the size of the NLR-surrounding genomic regions were biggest in the assembly of the resistant *S. americanum* accession that provided the longest input reads (see figure 1b in Witek et al. (2016b)). The number of predicted NLRs was higher in this dataset,

too. 401 full and 245 partial NLRs were found, whereas only 322 full and 293 partial NLRs were detected in the susceptible cultivar with smaller input fragment size. The assembly and NLR annotation differences should not solely be attributed to the input read length, but also to the total amount of bases in sequenced reads which differed by 43 Mb.

Similar results were reported for *Solanum verrucosum* (Giolai et al. 2016). They sequenced 2 SMRT cells for libraries with fragments centered around 4 kb, 5 kb, and 7 kb and assembled ROIs with `HGAP3` (Chin et al. 2013). Both assembly size and contiguity increased with increasing fragment size, and NLRs were more often reported to be complete. They also exemplified the capacity to bridge intergenic non-NLR regions with a 25 kb NLR cluster containing two NLR genes. Only the largest library provided ROIs that could be assembled to one contig containing both NLRs and the intergenic region.

Results from *S. americanum* and *S. verrucosum* can only be qualitatively transferred to the pan-NLR'ome study of *A. thaliana*. Both papers layed out advantages of longer input reads, but they did not take read quality into account, or the amount of reads that is needed for a reliable assembly.

In this project it was shown how input amounts (reads and total bases), read length, and read quality influenced the assembly size and NLR annotation capacity of four different *A. thaliana* accessions. Results showed that only sufficient ($\sim\geq$10 k) high quality ($\sim\geq$Q20) input reads secured a reliable and complete annotation of the NLR gene complements in those accessions.

Two accessions (Col-0 and Ty-1) represented good RenSeq experiments with many reads, whereas the two other accessions KBS-Mac74 and Cvi-0 showed less successful RenSeq experiments that resulted in fewer sequenced reads (fig. 2.2). The read quality distribution was similar for all four accessions, which at a first glance did not point to an additional problem next to a low read count for KBS-Mac74 and Cvi-0.

But already a critical evaluation of read lengths and read quality (fig. 2.3) showed additional differences. In the good RenSeq experiments, the read length distribution was independent from the read quality, with a mean read length around 2900 bp whereas the two less successfull experiments provided fewer long reads with high quality. This suggested a short polymerase read length (life time of the polymerase, see section 1.5.2) in KBS-Mac74 and Cvi-0. Only short fragments resulted in enough subreads for a high quality consensus call (fig. 1.6), whereas longer fragments were sequenced less often and produced CCS reads of minor quality.

On the one hand an assembly benefits from many input reads because the high coverage allows for a reliable base call per position. On the other hand, low-quality input reads might confuse the assembler, increasing fragmentation and preventing correct base calls. Long reads improve the contiguity of an assembly because they can span complex regions and regions with low coverage, but long reads often come with the above mentioned lower per base quality. The trade-off between high-quality reads, long reads, and the total amount of reads determines the assembly size, correctness, contiguity and accordingly the completeness of the NLR complement.

Evaluations of the assembly size (fig. 2.4), NLR numbers and NLR completeness (fig. 2.5), as well as misassembled positions (fig. 2.6) and mismatched bases (fig. 2.7)

led to the decision to discard any read with a quality below Q=20. Thus, all reads contained at most 1 % erroneous base calls and enough long reads remained to assemble mostly full NLR genes in the high quality datasets from Col-0 and Ty-1. For Cvi-0 and KBS-Mac74 the results clearly showed that these datasets did not provide enough reads of high quality and sufficient length to reliably assemble the accession-specific NLR complements. This knowledge was applied to the initial set of 73 *A. thaliana* accessions to filter out some low-quality datasets (section 2.1.3) that otherwise would have distorted the pan-NLR'ome analyses reported in chapter 3.

Testing the influence of input reads, read length, and read quality in four RenSeq datasets with (unpublished) reference genomes, provided useful information about the overall assembly quality and subsequent NLR annotation potential that a RenSeq dataset typically provides. Sufficient high quality input reads secured a reliable and complete annotation of the NLR gene complements in those accessions.

The analysis of the NLR complements of 65 accessions also showed later in the project that a high assembly quality and NLR annotation completeness was obtained (fig. 2.16 and fig. 3.A.13). The filtered SMRT RenSeq input enabled the assembly of full NLRs, NLR clusters (fig. 3.1b), and surrounding genomic regions. Regulatory elements in the ±3 kb neighborhood of NLRs were assembled. These might provide useful information for applied functional studies that for example need the promoter sequence of an NLR. Non-NLR genes in the flanking regions of NLRs were assembled, too. They can be used for example to anchor NLRs that are not present in Col-0 to the respective genomic position in the reference genome (section 3.4.5).

Without an accession-specific high quality reference genome (whose existence in turn would make RenSeq obsolete), quantitative evaluations of 'sufficient' and 'high Quality' input reads are not possible when starting a RenSeq project. A conservative advice would be to sequence at least one SMRT cell per *A. thaliana* accession. In this study, this resulted in ∼80 Mb of sequence data and together with read lengths ∼3 kb secured a reliable assembly and NLR annotation.

For studies in other species, I strongly suggest a test run using SMRT RenSeq in the reference accession. After running the whole assembly and annotation pipeline, results should be compared to the known NLRs from the reference genome to confirm that enough high quality reads (and bases) at the given read length were sequenced to annotate the full NLR complement.

Future methodological research could focus on how to reduce the amount of input needed. A possible starting point for tests could be increasing the fragment length even further which more easily leads to reliable overlaps during the assembly process. However, increasing the fragment length reduces the CCS read quality due to less passes per subread defined by the life of the polymerase (see section 1.5.2). Combining high quality CCS reads with CLR reads of lower quality but longer read length, would profit from both read quality and read length. Low-quality CLR reads were produced, but had to be discarded in this study. Finding a way to use them to improve the assembly continuity provides a cost-efficient (zero additional costs) and desirable outcome of future research. Long reads could also be created with Oxford Nanopore's long read sequencer MinION. The MinION is a small portable device (size of a USB thumb drive) that creates

read lengths equivalent to PacBio's SMRT CLRs with a similar error rate. Giolai et al. (2017) used Nanopore data to repeat the RenSeq experiment from Witek et al. (2016b) and showed similar results.

The assembly of long and erroneous reads is an ongoing field of study in bioinformatics and assemblers are continuously being developed and adjusted. A re-evaluation of possible assembly methods and parameters is thus always recommended.

4.2. WGS data for improved assemblies and assembly quality assessment

Independent of the species, accessions that are subjected to RenSeq studies have often already been sequenced in other projects before. Public databases store enormous amounts of whole genome sequencing short read data, and also the amount of long read data increases continuously. There is no reason why a RenSeq project should not make use of these datasets. Accession-specific WGS data are extremely useful for the assembly of RenSeq data, and provide additional possibilities to assess the assembly quality.

4.2.1. Using WGS to improve the assembly

The strength of RenSeq can also be its weakness. RenSeq is NLR-centered which reduces the assembly complexity and provides a very high coverage (several 100x possible, Andolfo et al. (2014), Jupe et al. (2013), and Witek et al. (2016b)) in NLR regions. But saying it differently, RenSeq suffers from low coverage in intergenic non-NLR regions which can complicate the correct assembly of big NLR clusters with long interspersed non-NLR regions. Long SMRT RenSeq reads (especially CLRs, but also CCS reads) should be able to span some of the intergenic non-NLR regions, but they are only present at low coverage and thus introduce erroneous bases to the assembly.

A concept that is regularly applied in whole genome assembly projects is to use short and nearly error-free reads to correct assemblies that were produced with long and more erroneous reads. The *A. thaliana* accession Nd-1 was assembled with PacBio reads (`FALCON` and `Canu`) and Illumina short reads were used to detect errors in the chondrome and plastome (`CLC` Basic Variant Detection from CLCbio (2018), Pucker et al. (2018)). The Durian assembly published by Teh et al. (2017), is a PacBio *de novo* assembly (`FALCON`, Chin et al. (2016)) corrected with short Illumina reads (`PILON`, Walker et al. (2014)). The maize reference genome was improved using PacBio reads and optical maps for the assembly (`Celera` and `FALCON`), and short reads for polishing (`PILON`) (Jiao et al. 2017)). Of course also non-plant genome assemblies profit from long read assemblies combined with short read polishing. The *de novo* assembly of a Korean human genome (Seo et al. 2016) was produced with PacBio reads and optical maps (`FALCON`), and polished with short Illumina reads (`PILON`).

Witek et al. (2016b) show that short Illumina reads from whole genome sequencing could be used to improve the per-base quality of the SMRT RenSeq assembly. The assembled C18 NLR cluster (276 kb) was corrected with short Illumina reads (HiSeq2000,

paired end (PE), 90bp to 30x coverage) together with 250 bp short reads from a RenSeq experiment. Strict filtering prevented using mismapped reads. They used only correctly mapped read pairs and examined only positions with a coverage between 20x-40x from WGS reads and 150x-500x from RenSeq reads. They then mapped back the ROIs that were used to assemble the cluster and reported a minimum per-base quality of 98.6 % (mean=99.6 %). Of the 171 found single nucleotide errors, more than 50 % were found in regions assembled from less than three ROIs, for example at contig ends. No other published RenSeq paper reported the use of WGS short reads to correct the assembly because intergenic regions were out of the scope of the analyses.

There are also other possibilities how an assembly can be created from the combination of long and short read data. For the *A. thaliana* accession Landsberg, short Illumina reads were used for assembly (`ALLPATH-LG`, Gnerre et al. (2011)) and scaffolding (`SSPACE-ShortRead`, Boetzer et al. (2011)) by Zapata et al. (2016). Long PacBio reads were here used only for gap-closing (`PBJelly`, English et al. (2012)). Applying this to a RenSeq-centered project is probably not advisable though, because the assembly would suffer from all the short-read derived problems introduced in section 1.4.2.

Instead, long PacBio CLRs could be corrected using short Illumina reads prior to a *de novo* assembly, for example using the program `proovread` (Hackl et al. 2014). This error correction step was for example applied in the assembly pipeline that created a reference genome for the European beech (Mishra et al. 2018), or the genome of the fungus *Agrocybe aegerita* (Gupta et al. 2018). For a RenSeq study, a possible scenario would be to *de novo* assemble RenSeq CCS reads together with corrected RenSeq CLRs, and to polish the assembly with short reads. This would not only provide a higher per base accuracy, but also extend the assembly further into neighboring non-NLR regions. One could also imagine to combine two *de novo* assemblies, one using only SMRT RenSeq reads, the other one using only short whole genome sequencing Illumina reads. A third alternative would be to assemble all datasets together, but the uneven coverage of the RenSeq data could be problematic since most assemblers expect even genomic coverage to work correctly.

Whole genome sequencing short read data for the 65 *A. thaliana* accessions were created only in a late stage of this project when the assembly pipeline had long been established. Thus, testing the different possibilities that come with the combination of whole genome sequencing short reads with SMRT RenSeq long reads need to be part of a different, bioinformatics-centered project with the advantage that many necessary datasets have already been produced. Here, the data is used for quality control only (section 4.2.2).

Last, I want to make the confrontative statement that for small genomes like *A. thaliana*, RenSeq no longer is the best option for a pan NLR'ome centered project. In the current era of cheap long read sequencing and even cheaper short read sequencing options, using whole genome long reads (PacBio SMRT or Oxford Nanopore) at best in combination with short read data for polishing might be the better choice. The sequencing cost will be higher than for RenSeq, but the data processing (quality control, read trimming, read correction, assembly) follows standard pipelines that have been established and applied in many research projects before. The analysis is thus less labor

and time demanding.

The advantages of a SMRT RenSeq assembly would be surpassed in a whole genome SMRT assembly. All NLR genes would be completely assembled, their surrounding regions with all regulatory elements would be present, and even the biggest, several hundred kb long NLR clusters could be analyzed. Accession-specific NLR'omes could then more easily be compared via whole genome alignments. Synteny would add another layer of information, for example guiding the orthogroup construction. The 'sixref' project is an ongoing study dedicated to produce six 'platinum' *A. thaliana* genomes based on SMRT technology and optical maps. The analysis presented in section 2.1.2 uses the yet unpublished reference genomes for Col-0, KBS-Mac74, Cvi-0, and Ty-1 and already provides proof of the superiority of these assemblies in comparison with the RenSeq assemblies. RenSeq assemblies were almost complete in the two successful RenSeq studies (Col-0 and Ty-1 in fig. 2.5), but the platinum whole genome assemblies contained the full NLR'omes (shown as dashed green lines in fig. 2.5). Another project dedicated to assemble a reference genome for the accession Nd-1 is already in the publication process (Pucker et al. 2018).

RenSeq with long SMRT PacBio reads is nonetheless an extremely useful technique with the advantage of being cheap and focused on NLRs. Especially when analyzing many and big genomes, heterozygous non-selfers and polyploid species, it is favorable over whole genome sequencing. Armstrong et al. (2018) for example compared dRenSeq with WGS reads in potato. They showed that 88x more WGS data (69.04 Gb) was needed to find the NLRs and intergenic regions they reported with dRenSeq using only one twelfth of a MiSeq flow cell (0.78 Gb).

The TwoBlades research effort which the *A. thaliana* NLR'ome project is embedded in, focuses on those more complex crop genomes (2Blades 2015). For these species, RenSeq is inevitable and the results and techniques reported in this thesis and soon-to-be published in the paper built on chapter 3 will provide useful information and guidance.

4.2.2. Quality Control (QC) with WGS read data

Critical method revisions and refinements, as well as quality control of the data are needed to secure the reliability of any biological analysis and the drawn conclusions. Unprocessed and thus genuine RenSeq input reads provide a useful source for the quality control (QC) of the Assembly Quality and Completeness of the created NLR complements (section 2.4). QC based solely on the RenSeq input sequences has the additional advantage that other RenSeq-based research projects can directly apply the here introduced methods, without the need to produce additional datasets.

But, next to being useful for the assembly of NLRs, WGS short read data can provide valuable support for the QC of NLR complements, and enable some analyses not possible with RenSeq data alone. The WGS reads form a second, unbiased dataset in addition to the enriched RenSeq dataset. They reflect the whole genome, not just the NLRs, come with equal coverage independent of genomic content and have a high read quality independent of the read length.

Whole genome sequencing reads were produced for all 65 *A. thaliana* accessions in parallel with the RenSeq data, using Illumina short read sequencing. DNA was extracted using the DNeasy Plant Mini Kit from QIAGEN. PCR-free libraries with a fragment size of 450 bp were created using the TruSeq DNA PCR-Free library preparation kit from Illumina and were sequenced on a MiSeq machine in paired end (PE) mode, with 250bp read length. Each accession was sequenced to ∼50 fold coverage.

The completeness analysis (section 2.4) benefits from using WGS reads instead of CCS reads because they come with a higher per-base accuracy and cover all NLRs equally (section 4.2.2.1). PCR-free WGS short read mappings can also be used to detect putative collapsed NLR gene clusters in the RenSeq assemblies (section 4.2.2.2).

4.2.2.1. Completeness analysis using WGS read data

WGS short reads can be used to calculate the Quality and the Completeness analogously to the CCS-read based completeness analysis (section 2.4). The short reads provide a very high per-base quality (∼ 99 %) and pseudo-heterozygous calls are more reliable than those created by CCS reads. If - in rare cases - NLRs escaped the enrichment, they would not be present in the RenSeq CCS reads and thus not create pseudo-heterozygous SNP calls at all. The Completeness would be overestimated in those cases. In contrast, a WGS experiment is not depending on an enrichment. All NLRs are present with equal coverage leading to a correct representation of an accession's Completeness.

I used PCR-free 250bp PE Illumina reads (450bp insert size, coverage=50) to calculate the Quality and the Completeness of the sub-sampled Col-0 assemblies and the RenSeq accessions. Reads were trimmed (`Skewer`; version 0.1.124; -Q30 -q30 -l36; Jiang et al. (2014)), mapped to the respective pseudogenomes (`BWA` mem; version 0.7.17-r1194-dirty; Li et al. (2009a)), and only high quality mappings were used (`SAMtools`; version 1.9; Li et al. (2009b); samtools view -q60 -f2 -F2052; samtools rmdup). Quality and Completeness were calculated as described in section 2.4.

The generally high Completeness of my RenSeq datasets was confirmed by the short read based analysis (median=95 %). Compared to the CCS-read based results, the Completeness showed a slight decrease and a wider range.

41 accessions (compared to 46 accessions in the CCS-read based analysis) were at least 95 % complete (vertical black line in fig. 4.2.1 panel b), and 49 accessions (compared to 56) were at least 90% complete. Completeness values ranged between 12 % and 97 % (63 %-97 % CCS-read based). The accession with only 12 % completeness (7308, Po-0) was an outlier and the next accession already showed a 56 % complete NLR complement. 14 accessions had a higher Quality than the full test Col-0 dataset and could not be ranked. Their completeness (unfilled circles in fig. 4.2.1 panel b) was between 97 % and 100 %.

The Completeness calculation method is not perfect, which was reflected in the Col-0 RenSeq dataset which showed the highest Quality and Completeness of all accessions, and exceeded the results of the corresponding full test Col-0 dataset. This can be explained by slight annotation and mapping differences. NLRs had to be defined based on exonerate mappings for the sub-sampling experiments because no manually curated NLR

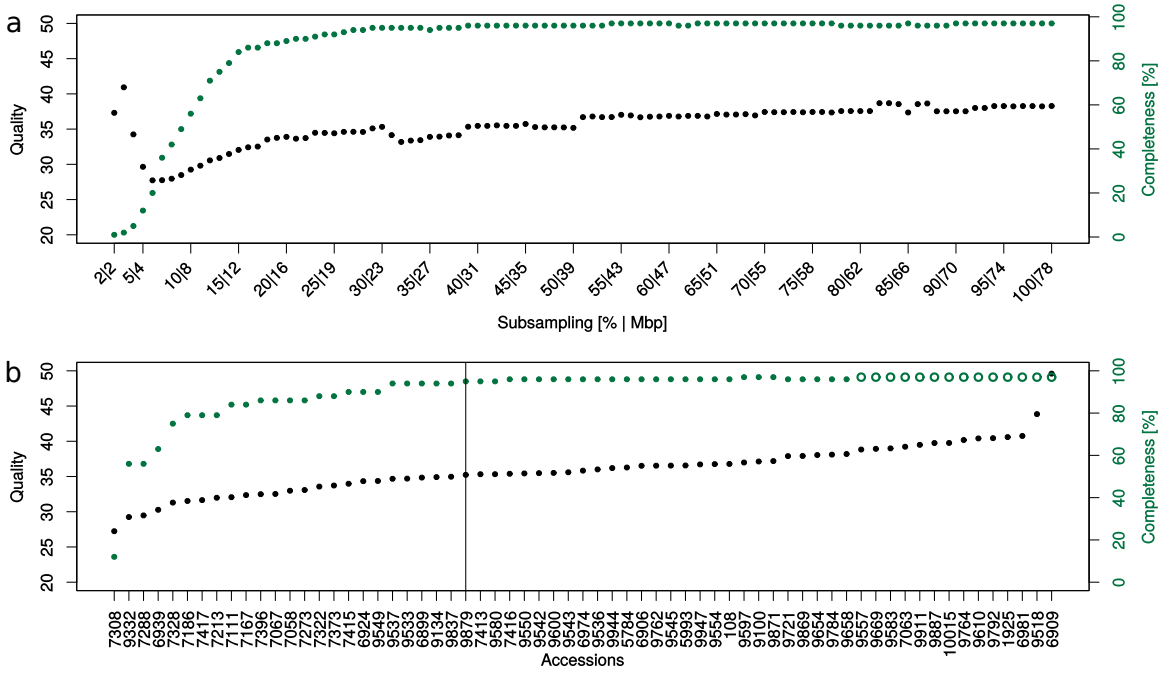


Figure 4.2.1.: Short read based Quality and Completeness
a) Quality (black) and Completeness values (dark green) for sub-sampled Col-0 datasets. The x axis contains the sub-sampling percentage and the amount of input data for each sub-sampling experiment. b) Quality (black) and Completeness values (dark green) for all RenSeq accessions. Unfilled circles indicate accessions with qualities larger than any sub-sampled dataset. The vertical black line is drawn at 95 % Completeness.

annotations existed. Small annotation differences remained and could not be improved further. Mapping heuristics also influenced the Completeness results. Even extensive parameter optimization did not resolve every mapping difference. Only accession-specific reference genomes could provide perfect Completeness values, but in turn would make RenSeq obsolete.

4.2.2.2. Detection of assembly errors: collapsed NLRs

Clustered NLRs with highly repetitive coding sequences can be hard to assemble correctly. They might be erroneously collapsed by the assembler, especially if they are recent tandem duplicates which sit next to each other in the genome and share similar surrounding regions with low information content (e.g TEs). Long SMRT RenSeq reads normally prevent collapsing NLR genes during the assembly, but this claim needs to be confirmed. PCR-free short read mappings can be used to detect collapsed NLRs in RenSeq assemblies by searching for excessively highly covered NLR genes. PCR-free reads are needed because only those provide an even coverage of the genome. Libraries prepared with PCR, would be prone to amplification bias and thus not be usable for a coverage-based detection of collapsed NLRs. I searched for collapsed NLRs in the 65 assemblies using accession-specific PCR-free MiSeq reads. The read mapping procedure was already described in section 4.2.2.1. The expected mean coverage of each accession was defined using only exons from the benchmarking set of universal single copy ortholog (BUSCO) genes (Waterhouse et al. 2013). This conservative method excluded coverage fluctuations in paralogous genes that might come from an uneven distribution of read mappings. It also prevented deleterious effects that might come from intron mappings. Introns sometimes contain TEs or other regions of low information content that attract erroneous read mappings which would artificially inflate the mean coverage of a gene. The mean NLR coverage was calculated accordingly, also only using exons, and was normalized with the accession-specific mean coverage of the BUSCO genes. Correctly assembled NLRs were expected to have a normalized coverage around one. Outliers with high coverage were candidates for a collapsed NLR cluster.

The method suggested high quality assemblies without many collapsed NLRs (fig. 4.2.2 and table A.4). Only 51 genes were reported with a normalized coverage $\geq$2. Of those, 23 NLRs contained only a RPW8 domain, suggesting that RNL clusters were generally hard to assemble. The two NLRs with the highest normalized coverage were the neighboring RNLs *7396|T234-R1* and *7396|T235-R1* (accession Ws-0). They were found on a 10kb long contig in a cluster with three other RNLs. These RNLs were within the top 12 of the putative collapsed NLRs (table A.4), further supporting a collapsed RNL cluster.

My method to detect collapsed NLRs was developed to be conservative and specific. Strict filtering of mapped reads was applied to exclude mappings from genomic regions with similar sequence content. These mappings would artificially have inflated the coverage and the number of putative collapsed NLRs. This behavior was necessary to detect collapsed NLRs, but led to an artificial increase of lowly covered NLRs especially at contig ends. Lowly covered NLRs (left peak in fig. 4.2.2) can thus not be interpreted and are not a hint for inflated NLR clusters.

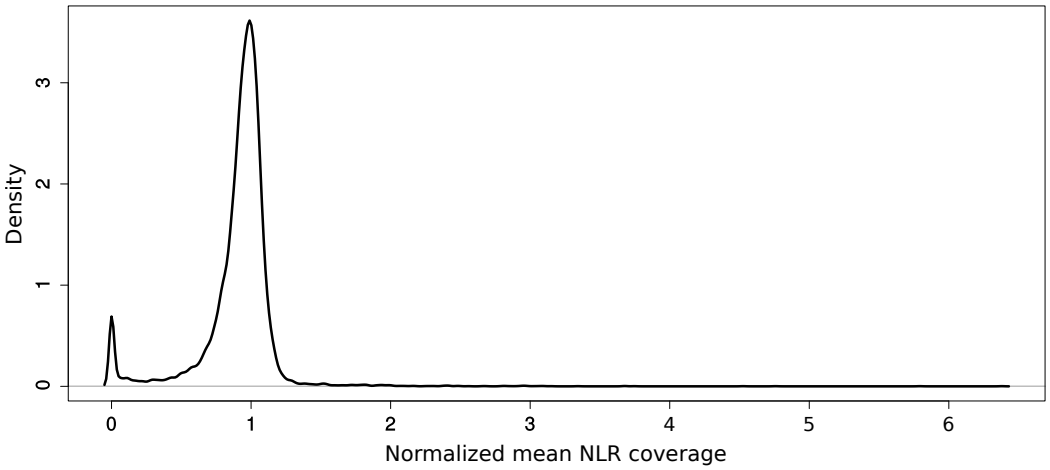


Figure 4.2.2.: Normalized NLR coverage distribution
Density distribution includes all NLRs from all accessions.

4.3. Gene annotation

I showed the major gene annotation problems for NLR genes and how the automated gene annotation was iteratively optimized to remove as many errors as possible. Manual curation was THE key to secure the highest possible correctness for NLR gene annotations, a notion that is widely accepted in the scientific community (Yandell et al. 2012) and led to big 'Annotation Jamborees', collaborative efforts to manually improve the correctness of automated gene annotations. Genes of *Drosophila* (Hartl 2000), bees (Elsik et al. 2006) and ants (Munoz-Torres et al. 2011) have for example been manually curated in such jamborees.

The long-term goal is to develop automated gene annotation pipelines that make manual curation expendable. This still requires lots of bioinformatics research and is possibly not within reach in the next years. In the meanwhile, there are several options to at least improve the automated annotation. In addition to the ab initio gene predictors `AUGUSTUS` and `SNAP`, `MAKER` comes with `GeneMark-ES` (Lomsadze et al. 2005) and `Fgenesh` (Solovyev et al. 2006). Using more predictors increases the probability to find the correct gene model. Especially when supported by evidence, it will be easier for `MAKER` to select the best fitting prediction and combine it with the evidence to create the optimal gene annotation.

Gene predictors are using predefined settings of organism-specific traits like typical intron- and exon lengths or codon frequencies (Yandell et al. 2012). Training the gene predictor with organism-specific data is typically advised for non-model species. Here, `SNAP` was trained with NLR data to produce the HMM to use for gene prediction. `AUGUSTUS` could also be trained with NLRs instead of using the internal *Arabidopsis*-specific profile. NLR-specific profiles might have the downside of being less specific on the neighboring non-NLRs, but this would not be as grave as incorrect NLRs. `MAKER` sup-

ports iterative gene annotation, using the previous genes as training sets for subsequent annotation runs. This could improve the NLR annotation further.

The *ab initio* predicted genes are improved by `MAKER` using additional evidence. Transcripts, expressed sequence tags (ESTs), and proteins from the reference accession Col-0 were used to support the correct annotation of intron- and exon boundaries (section 2.2.1). Accession-specific transcript evidence would have the potential to further improve automated gene annotations. RNA-Seq data could be assembled *de novo* with `Trinity` (Grabherr et al. 2013) or mapping-based with `TopHat` and `Cufflinks` (Trapnell et al. 2012), providing exact exon- and intron structures. Long read sequencing methods like PacBio's Iso-Seq provide full-length transcripts that would not even have to be assembled, removing possible errors during that step.

Accession-specific NLR transcripts were produced with cDNA RenSeq (Illumina short reads) in Witek et al. (2016b). They were used to detect false-positive frame-shift mutations and expressed NLRs. Such datasets could also be directly used in the automated annotation pipeline. The full set of NLR transcripts will hardly be producible because NLRs are often only expressed at very low frequencies, or only upon pathogen induction. Even though WGS based transcripts will contain less NLR transcripts than the enriched cDNA RenSeq datasets, they could prevent the fusion of non-NLRs to NLRs. Non-NLRs would be supported by enough transcript evidence to be correctly annotated, which would allow `MAKER` to choose the correct gene model for the neighboring unexpressed NLR gene, too.

The manually curated NLRs of this project can be used as evidence for other *A. thaliana* RenSeq projects and also for closely related species. `MAKER` uses evidence from related species as 'alternative evidence', giving less weight to those compared to intraspecific evidence lines.

There are other tools that can be used in addition to `MAKER` to improve gene annotations. The `EVidenceModeler` (Haas et al. 2008) is a gene annotation tool that creates genes from the weighted consensus of all used evidence. It is capable to report alternative splice variants, whereas `MAKER` only reports one representative isoform per gene. The tool `deFusion` (Wang 2018) was written to fix false gene merges in `MAKER` gene annotations, caused by evidence that bridges neighboring paralogous genes. This was one of the problems that resulted in many of the NLR gene fusions detected in this study (section 2.2.1.1). In addition, `deFusion` also detects fusions caused for example by wrongly assembled transcripts from RNASeq data, expanding its applicability beyond the detection of wrong evidence mappings.

Still, there is currently no way around manual reannotation to secure a high accuracy of the predicted gene structures. The development of a standard operating procedure (SOP) for *A. thaliana* NLR reannotation (section 3.C.1) was an important mile stone of this project that can be adopted for other RenSeq studies independent of the species. The annotation jamboree of this project was a collaborative effort of four skilled researchers with profound experience with NLR gene architectures in Brassicaceae (Freddy Monteiro and Oliver Furzer), with knowledge of the bioinformatics background, the features and the problems of automated gene annotation (Felix Bemm and Anna-Lena Van de Weyer), and with experience in the manual curation of genes (Felix Bemm). The developed

SOP contains all necessary basics to find and to reannotate false gene predictions. It also suggests various flags that can be used to rate the reliability of a (re)annotated gene. The flag 'corbound' for example marks genes with changes in exon- and intron-boundaries that are not unequivocally supported by transcript or protein evidence. In addition, the flag 'cortrans' was used if the reannotation changed the initial translated region. Often, those changes were justified by rescuing the translation of known NLR domains. One has to keep in mind though, that this could also introduce the artificial translation of an actually pseudogenized gene. Thus, 'cortrans' and 'corbound' denote reannotations to be taken with care.

Truncated NLRs at contig borders were flagged with 'truncated' if protein or transcript evidence suggested a longer ORF. This allows to quickly identify putative incomplete gene annotations. Orthogroups and their phylogenies are based on protein alignments. Truncated genes lack certain positions in those alignments and, depending on the threshold that is applied, might prevent those positions to be considered for the alignment. If for example all of the NLRs need to cover a certain position for it to be included in the alignment, even a highly informative canonical domain could be excluded due to only one truncated gene. When thoroughly analyzing the variation of genes in a specific orthogroup, unflagged truncated genes suggest more variation than actually correct, could change the phylogeny of the orthogroup, or even artificially attract other paralogous genes. Referring back to section 4.2.1, the number of truncated NLR genes would go down if methods are developed that increase the amount of assembled flanking regions.

Follow up projects in *A. thaliana* or other species would profit from a pipeline that checks for putative annotation errors and points to problematic NLR genes that could be prioritized for manual curation. `MAKER` provides two quality measurements for each gene annotation. The Annotation Edit Distance (AED) reports how good an annotation fits the underlying evidence (Eilbeck et al. 2009). The AED lies between zero and one, zero meaning the perfect fit between evidence and gene annotation. The pipeline could for example report genes with an AED near one for prioritized curation. The second quality measurement is the quality index (QI) score, that summarizes in more detail features of annotated genes and how well the annotated transcripts fit the underlying data (Campbell et al. 2014). The QI score contains nine values (table 4.3.1) from which for example the quality of predicted splice sites and exons can be inferred. Using the QI score to further prioritize gene reannotation would be easy to implement.

In a RenSeq based assembly, one expects most NLR genes to contain canonical domains (NB, TIR, CC, LRR, and RPW8). The most prevalent NLR classes in many species are TNLs and CNLs. A TNL contains the domain sequence TIR-NB-LRR and CNLs contain CC-NB-LRR. The pipeline could check for each NLR if it deviates from these domain architectures, and if so, check if the neighboring gene contains the expected missing domain(s). This would point to a gene splitting event during automated annotation. Further evidence can be drawn from canonical domains found in the corresponding *ab initio* gene predictions. If these contain canonical domains not present in the final annotated gene, reannotation should be prioritized, too.

Genes containing IDs should be curated manually to make sure no annotation error led to an artificially fused novel domain. If researcher time is limiting, I suggest to prioritize

Table 4.3.1.: Quality index (QI) score
Description of the QI score produced by `MAKER` for each annotated gene.

column	definition
1	5' UTR length [bp]
2	Splice sites confirmed by EST/mRNA-seq alignments (fraction)
3	Exons matching EST/mRNA-seq alignments (fraction)
4	Exons overlapping with aligned EST/mRNA-seq or proteins (fraction)
5	Splice sites confirmed by ab initio gene predictors (fraction)
6	Exons overlapping an ab initio prediction (fraction)
7	Number of exons
8	3' UTR length [bp]
9	Protein length [bp]

ID containing gene models above all others for manual curation. The prioritization pipeline could directly report ID-containing genes. Finally, genes close to contig borders should be reported by the pipeline, because these could putatively be truncated genes that need to be flagged.

If, despite of all guidance and advice given here, manual annotation is not desired or not feasible, researchers need to accept the consequences. Fusion genes will artificially blow up the architectural diversity of the NLR'ome and incorrect IDs will be found. Depending on the quality of used evidence lines, these numbers can be extremely high. Even in the model organism *A. thaliana* with a very high amount of manually curated protein and transcript evidences and with species-centered *ab initio* models, 736 incorrect gene fusion (∼11 per accession) were found and corrected manually. Artificially split genes will increase the number of NLRs seeming to be biological truncations, which are known to be functional in some cases (Nishimura et al. 2017; Xiao et al. 2001). Truncated genes at contig ends will come with the orthogroup-related problems mentioned above (reduce the number of usable positions for the alignment, suggest higher variation, induce strange phylogenies, and attract paralogs). The orthogroup refinement method (section 3.7.3.2) should remove wrongly annotated genes from the refined orthogroups, which is of course desired, but would blow up the number of singleton genes in the cloud, and would reduce shell and core orthogroups suggesting more diversity than appropriate. I expect that these errors have had an influence also on already published NLR studies relying solely on automated gene annotations. How much the results would change after manual reannotation would need to be evaluated case by case.

4.4. Current and future use of the data

The pan-NLR'ome project in *A. thaliana* provided an in-depth view on the intraspecific diversity of NLR genes using 65 diverse accessions (presented in chapter 3). The data that was produced unlocks the potential for many more analyses. Research projects

focusing on specific NLR genes profit for example from the extensive orthogroup information that allows them to compare known functional sites in the different accessions.

Contigs and CCS reads have been published already on October 15th 2017 on the TwoBlades web page (TwoBlades Foundation 2017) to allow the public to advance their own studies of the plant immune system. All long RenSeq PacBio reads, the short whole genome sequencing Illumina reads, and the contigs will be published at the European Nucleotide Archive (ENA) upon publication of the manuscript. All gene annotations (gff files), and the domain positions of all NLR genes are already prepared for publishing via GitHub (currently not publicly accessible), and will be made public together with the manuscript (chapter 3). GitHub will also host all metadata information that was collected for accessions, transcripts, domains, architectures, and orthogroups (OGs). Orthogroup lists will be published there, too, together with results from architecture- or orthogroup-related analyses. These extensive resources will not only enable others to evaluate the analyses that were done in the pan-NLR'ome project, but also to carry on with NLR-focused research to further broaden our understanding of plant immunity.

All `MAKER` annotations and re-annotated NLRs together with all the used evidence is currently hosted at a private instance of `WebApollo` (see section 2.2.2). All will be made public when the paper is published. This way, others can reproduce the reannotation, include additional evidence lines and (locally) change existing annotations if there is enough persuasive evidence. More accessions can be easily incorporated if additional RenSeq experiments are carried out in the future.

Many research projects are centered around the function of one or few NLR genes. People therefore profit from getting easy access to the orthogroup that contains the NLR of interest and shows its presence in the different accessions. All original unrefined orthogroups and their phylogenies are already prepared for publishing at `iTOL` (Letunic et al. 2016). Users will be able to view the underlying alignments to find positions they are interested in. In addition, `iTOL` will contain the refined orthogroup identifiers to directly illustrate the separation of overclustered orthogroups. The domain content of each protein will show the different architectures that are represented in an orthogroup. Other useful metadata information like the geographic origin, the population structure, and expression likelihood will be presented there, too.

An elaborate way to visualize the NLR'ome could be implemented using the program `panX` (Ding et al. 2017). It enables interactive exploration of the data by combining alignments and phylogenies with statistical charts, gene cluster tables, and metadata information. It would provide a perfect round-up of the results from `WebApollo`,GitHub, and `iTOL`.

The RenSeq datasets created here form a great resource for the *A. thaliana*-centered research of plant immunity. In a recently published paper about the 'Modulation of *ACD6* dependent hyperimmunity by natural alleles of an *A. thaliana* NLR resistance gene' (Zhu et al. 2018), data from the pan-NLR'ome project allowed for correlating the allelic variation of the *Suppressor of NPR1-1, Constitutive 1 (SNC1)* gene with the severeness of the autoimmunity phenotype caused by *Accelerated Cell Death 6 (ACD6)*.

The study focused on ACD6, a transmembrane protein that can positively regulate immune responses via the salicylic acid (SA) pathway (Lu 2003). A hyperactive allele

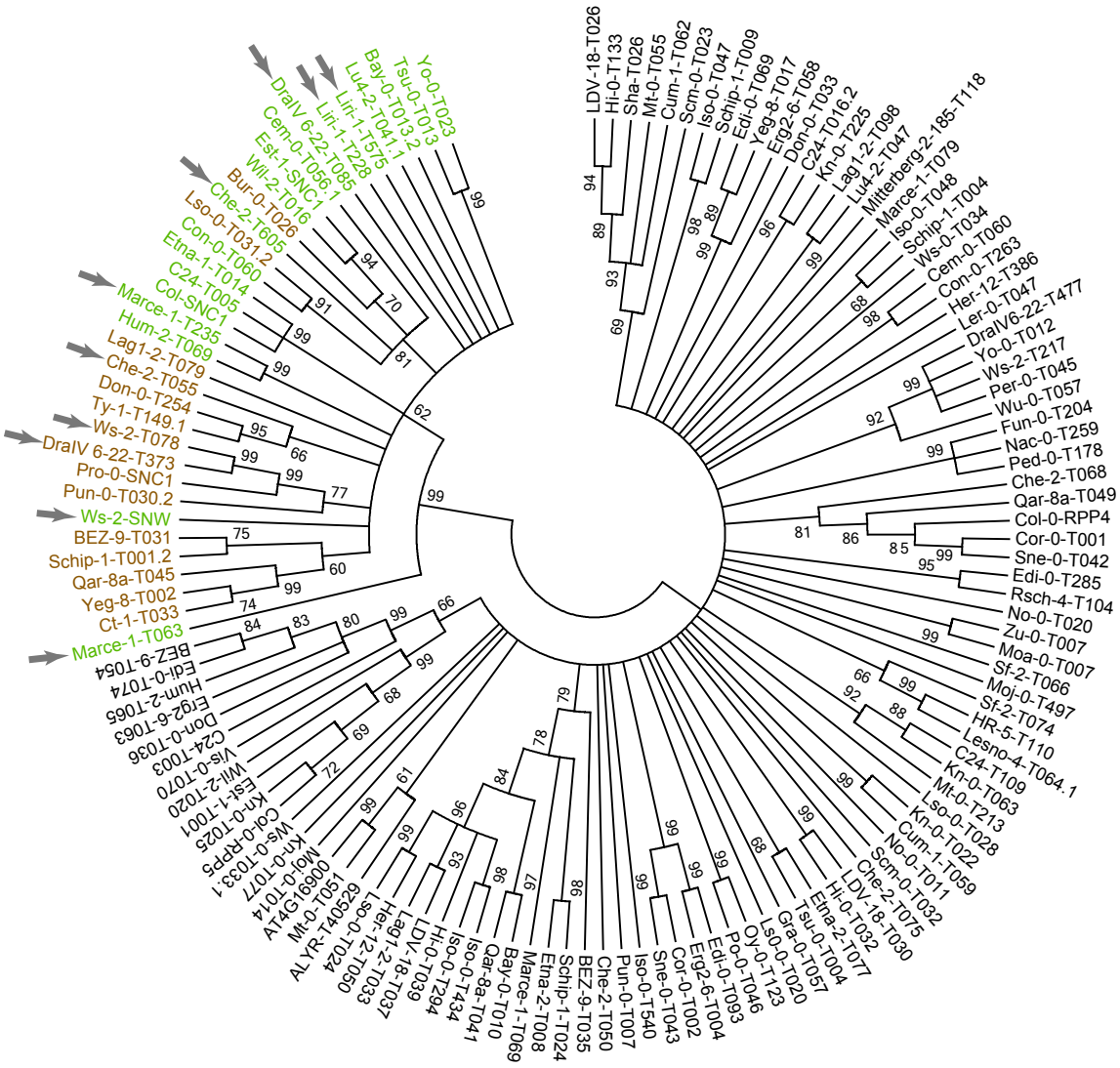


Figure 4.4.1.: Phylogeny of 136 RPP4/5 and SNC1 proteins from 65 accessions Bootstrap values over 60 % are indicated. The SNC1 clade is highlighted in color, with single NL linker (SNC1-sNL) genes green and duplicated NL linker (SNC1-dNL) genes brown. Arrows highlight five accessions with two SNC1 homologs. Reprinted with permission from Zhu et al. (2018) (Fig. 4A).

of *ACD6* was found in the natural *A. thaliana* accession Est-1 (Todesco et al. 2010). It was shown to increase resistance against several bacteria, fungi, and oomycetes, but at the same time triggers autoimmunity in absence of a pathogenic thread, which reduces the plant's growth and seed production. The allele was found in ∼20 % of 96 *A. thaliana* accessions (Todesco et al. 2010). 12 % of 823 accessions from the 1001G project (1001_Genomes_Consortium 2016) were shown to have a causal codon for the hyperactive *ACD6* allele (Zhu et al. 2018), further supporting the claim that the hyperactive *ACD6* is actively being kept in the population despite its potential fitness costs in nature.

Not all accessions with the hyperactive allele had the same severeness of autoimmunity, the accession Pro-0 for example did not show expected severe necrotic lesions. Zhu et al. (2018) showed that the activity of the hyperactive ACD6 allele can be modulated by several genes from other genomic loci. One of these loci was the *RPP4/RPP5* NLR cluster in genomic proximity (∼1Mb) to *ACD6*, containing *SNC1*, a known regulator of plant autoimmunity (Gou et al. 2012; Li et al. 2010b; Yi et al. 2007). Transgene approaches indeed suggested that the transcribed portion of SNC1 can attenuate ACD6 hyperactivity in Pro-0 background, leading the authors to the question how common different SNC1 alleles are in the *A. thaliana* population.

In the initial orthogroup sets created in the pan-NLR'ome project, the Col-0 proteins RPP4 and RPP5, and another, uncharacterized protein (AT4G16920) were placed into an overclustered orthogroup with many paralogs (OG4). SNC1 was in a smaller OG without other Co-0 paralogs (OG216). Due to the high sequence similarity between RPP4,RPP5, and SNC1, a protein-phylogeny was created from all these sequences together with the known SNC1 proteins from Est-1, Pro-0, and Ws-2, the known RPP5 from Ler-0, and the RPP4/RPP5/SNC1 homolog AT4G16900 from Col-0 (fig. 4.4.1).

The SNC1 clade was manually defined (Zhu et al. 2018) and indicated 34 *SNC1* genes from 29 accessions. The RenSeq data allowed to compare the sequences of different *SNC1* alleles and the most prominent difference was found in the NL linker sequence, which was duplicated in nearly half of the SNC1 proteins. This polymorphism in the NL linker was then experimentally shown to explain differences in SNC1 activity and also attenuation of ACD6 activity in Pro-0. RenSeq data from the pan-NLR'ome project nicely supported the main claims of the paper, and broadened the understanding of the intraspecific diversity of the evaluated causal *SNC1* locus.

For sure, other projects focusing on specific NLRs will profit in similar ways from the *A. thaliana* pan-NLR'ome datasets.

Bibliography

1001_Genomes_Consortium (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". In: *Cell* 166.2. URL: http://dx.doi.org/10.1016/j.cell.2016.05.063.

2Blades (2015). *Resistance Gene Diversity: Defining the building blocks of the plant immune system*. URL: http://2blades.org/projects-and-technology/projects/2/ (visited on 09/13/2018).

Ahuja, Ishita, Ralph Kissen, and Atle M. Bones (2012). "Phytoalexins in defense against pathogens". In: *Trends in Plant Science* 17.2. arXiv: /linkinghub.elsevier.com/retrieve/pii/S0960982205000989 [http:]. URL: http://dx.doi.org/10.1016/j.tplants.2011.11.002.

Albert, Isabell, Hannah Böhm, Markus Albert, Christina E. Feiler, Julia Imkampe, Niklas Wallmeroth, Caterina Brancato, Tom M. Raaymakers, Stan Oome, Heqiao Zhang, Elzbieta Krol, Christopher Grefen, Andrea A. Gust, Jijie Chai, Rainer Hedrich, Guido Van Den Ackerveken, and Thorsten Nürnberger (2015). "An RLP23-SOBIR1-BAK1 complex mediates NLP-triggered immunity". In: *Nature Plants* 1.

Albert, Markus (2013). "Peptides as triggers of plant defence". In: *Journal of Experimental Botany* 64.17.

Andolfo, Giuseppe, Florian Jupe, Kamil Witek, Graham J Etherington, Maria R Ercolano, and Jonathan D G Jones (Jan. 2014). "Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq." In: *BMC plant biology* 14.1. URL: http://www.biomedcentral.com/1471-2229/14/120.

ARAPORT: Arabidopsis Information Portal (2018). URL: https://www.araport.org/downloads (visited on 11/08/2018).

Araport11 JBrowse (2018). URL: https://apps.araport.org/jbrowse/?data=arabidopsis%7B%5C&%7Dloc=Chr4%7B%5C%%7D3A5970190..5975429%7B%5C&%7Dtracks=TAIR10%7B%5C_%7Dgenome%7B%5C%%7D2CAraport11%7B%5C_%7DLoci%7B%5C%%7D2CAraport11%7B%5C_%7Dgene%7B%5C_%7Dmodels%7B%5C%%7D2Clight%7B%5C_%7Dtophat%7B%5C%%7D2CTAIR9%7B%5C_%7DtDNAs%7B%5C%%7D2CSALK%7B%5C_%7DtDNAs%7B%5C&%7Dhighlight= (visited on 11/08/2018).

Armstrong, Miles R, Jack Vossen, Tze Yin Lim, B Hutten Ronald C, Jianfei Xu, Shona M Strachan, Brian Harrower, Nicolas Champouret, Eleanor M Gilroy, and Ingo Hein (2018). "Tracking disease resistance deployment in potato breeding by enrichment sequencing". In: *Plant Biotechnology Journal*. URL: https://www.biorxiv.org/content/early/2018/07/02/360644.

Bailey, Paul C., Christian Schudoma, William Jackson, Erin Baggs, Gulay Dagdas, Wilfried Haerty, Matthew Moscou, and Ksenia V. Krasileva (2018). "Dominant integra-

tion locus drives continuous diversification of plant immune receptors with exogenous domain fusions". In: *Genome Biology* 19.1.

Bakker, Erica G, Christopher Toomajian, Martin Kreitman, and Joy Bergelson (Aug. 2006). "A genome-wide survey of R gene polymorphisms in Arabidopsis." In: *The Plant cell* 18.8. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1533970%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Bao, Jiandong, Meilian Chen, Zhenhui Zhong, Wei Tang, Lianyu Lin, Xingtan Zhang, Haolang Jiang, Deyu Zhang, Chenyong Miao, Haibao Tang, Jisen Zhang, Guodong Lu, Ray Ming, Justice Norvienyeku, Baohua Wang, and Zonghua Wang (2017). "PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in Magnaporthe oryzae". In: *Molecular Plant* 10.11.

Bendahmane, Abdelhafid, Garry Farnham, Peter Moffett, and David C. Baulcombe (2002). "Constitutive gain-of-function mutants in a nucleotide binding site-leucine rich repeat protein encoded at the Rx locus of potato". In: *Plant Journal* 32.2.

Berardini, Tanya Z., Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily Strait, and Eva Huala (2015). "The Arabidopsis Information Resource: Making and Mining the 'Gold Standard' Annotated Reference Plant Genome". In: *Genesis* 53.8. arXiv: 15334406.

Bernoux, Maud, Thomas Ve, Simon Williams, Christopher Warren, Danny Hatters, Eugene Valkov, Xiaoxiao Zhang, Jeffrey G. Ellis, Bostjan Kobe, and Peter N. Dodds (2011). "Structural and functional analysis of a plant resistance protein TIR domain reveals interfaces for self-association, signaling, and autoregulation". In: *Cell Host and Microbe* 9.3.

Bigeard, Jean, Jean Colcombet, and Heribert Hirt (2015). "Signaling mechanisms in pattern-triggered immunity (PTI)". In: *Molecular Plant* 8.4. URL: http://dx.doi.org/10.1016/j.molp.2014.12.022.

Boetzer, Marten, Christiaan V. Henkel, Hans J. Jansen, Derek Butler, and Walter Pirovano (2011). "Scaffolding pre-assembled contigs using SSPACE". In: *Bioinformatics* 27.4. arXiv: NIHMS150003.

Böhm, Hannah, Isabell Albert, Stan Oome, Tom M. Raaymakers, Guido Van den Ackerveken, and Thorsten Nürnberger (2014). "A Conserved Peptide Pattern from a Widespread Microbial Virulence Factor Triggers Pattern-Induced Immunity in Arabidopsis". In: *PLoS Pathogens* 10.11.

Boller, Thomas and Georg Felix (Jan. 2009). "A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors." en. In: *Annual review of plant biology* 60. URL: http://www.annualreviews.org/doi/full/10.1146/annurev.arplant.57.032905.105346.

Bonardi, V., S. Tang, A. Stallmann, M. Roberts, K. Cherkis, and Jeffery L. Dangl (2011). "Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors". In: *Proceedings of the National Academy of Sciences* 108.39. arXiv: arXiv:1408.1149. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1113726108.

Brunner, Frédéric, Sabine Rosahl, Justin Lee, Jason J Rudd, Carola Geiler, Sakari Kauppinen, Grethe Rasmussen, Dierk Scheel, and Thorsten Nürnberger (2002). "Pep-13, a plant defense-inducing pathogen-associated pattern from Phytophthora transglutaminases". In: *The EMBO journal* 21.24.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, Vladimir Suvorov, and Andrey D. Prjibelski (2016). "RnaQUAST: A quality assessment tool for de novo transcriptome assemblies". In: *Bioinformatics* 32.14.

Campbell, Michael S., Carson Holt, Barry Moore, and Mark Yandell (2014). "Genome Annotation and Curation Using MAKER and MAKER-P". In: *Current Protocols in Bioinformatics*. arXiv: NIHMS150003.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, Xi Wang, Felix Ott, Jonas Müller, Carlos Alonso-Blanco, Karsten Borgwardt, Karl J Schmid, and Detlef Weigel (Oct. 2011). "Whole-genome sequencing of multiple Arabidopsis thaliana populations." In: *Nature genetics* 43.10. URL: http://dx.doi.org/10.1038/ng.911.

Cesari, Stella, Maud Bernoux, Philippe Moncuquet, Thomas Kroj, and Peter N. Dodds (2014). "A novel conserved mechanism for plant NLR protein pairs: the "integrated decoy" hypothesis". In: *Frontiers in Plant Science* 5.November. URL: http://journal.frontiersin.org/article/10.3389/fpls.2014.00606/abstract.

Cesari, Stella, G. Thilliez, C. Ribot, V. Chalvon, C. Michel, A. Jauneau, S. Rivas, L. Alaux, H. Kanzaki, Y. Okuyama, J.-B. Morel, E. Fournier, D. Tharreau, R. Terauchi, and T. Kroj (2013). "The Rice Resistance Protein Pair RGA4/RGA5 Recognizes the Magnaporthe oryzae Effectors AVR-Pia and AVR1-CO39 by Direct Binding". In: *The Plant Cell* 25.4. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.112.107201.

Chae, Eunyoung, Kirsten Bomblies, Sang-Tae Kim, Darya Karelina, Maricris Zaidem, Stephan Ossowski, Carmen Martín-Pizarro, Roosa A.E. Laitinen, Beth A. Rowan, Hezi Tenenboim, Sarah Lechner, Monika Demar, Anette Habring-Müller, Christa Lanz, Gunnar Rätsch, and Detlef Weigel (Nov. 2014). "Species-wide Genetic Incompatibility Analysis Identifies Immune Genes as Hot Spots of Deleterious Epistasis". In: *Cell*. URL: http://www.sciencedirect.com/science/article/pii/S0092867414013762.

Chen, Li Qing, Bi Huei Hou, Sylvie Lalonde, Hitomi Takanaga, Mara L. Hartung, Xiao Qing Qu, Woei Jiun Guo, Jung Gun Kim, William Underwood, Bhavna Chaudhuri, Diane Chermak, Ginny Antony, Frank F. White, Shauna C. Somerville, Mary Beth Mudgett, and Wolf B. Frommer (2010a). "Sugar transporters for intercellular exchange and nutrition of pathogens". In: *Nature* 468.7323. URL: http://dx.doi.org/10.1038/nature09606.

Chen, Qihan, Zhaoxue Han, Haiyang Jiang, Dacheng Tian, and Sihai Yang (2010b). "Strong Positive Selection Drives Rapid Diversification of R-Genes in Arabidopsis Relatives". In: *Journal of Molecular Evolution* 70.2. URL: http://link.springer.com/10.1007/s00239-009-9316-4.

Chen, Xinwei, Dominika Lewandowska, Miles R. Armstrong, Katie Baker, Tze-Yin Lim, Micha Bayer, Brian Harrower, Karen McLean, Florian Jupe, Kamil Witek, Alison K.

Lees, Jonathan D G Jones, Glenn J. Bryan, and Ingo Hein (2018). "Identification and rapid mapping of a gene conferring broad-spectrum late blight resistance in the diploid potato species Solanum verrucosum through DNA capture technologies". In: *Theoretical and Applied Genetics* 131.6. URL: http://link.springer.com/10.1007/s00122-018-3078-6.

Chen, Yu, Zhenyu Liu, and Dennis A. Halterman (2012). "Molecular determinants of resistance activation and suppression by phytophthora infestans effector IPI-O". In: *PLoS Pathogens* 8.3.

Cheng, Chia Yi, Vivek Krishnakumar, Agnes P. Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher D. Town (2017). "Araport11: a complete reannotation of the Arabidopsis thaliana reference genome". In: *Plant Journal* 89.4.

Chin, Chen Shan, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E. Eichler, Stephen W. Turner, and Jonas Korlach (2013). "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data". In: *Nature Methods* 10.6.

Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R Cramer, Massimo Delledonne, Chongyuan Luo, Joseph R Ecker, Dario Cantu, David R Rank, and Michael C Schatz (Oct. 2016). "Phased diploid genome assembly with single-molecule real-time sequencing". In: *Nature Methods*. URL: http://www.nature.com/doifinder/10.1038/nmeth.4035.

Choi, Kyuha, Carsten Reinhard, Heïdi Serra, Piotr A Ziolkowski, Charles J. Underwood, Xiaohui Zhao, Thomas J. Hardcastle, Nataliya E. Yelina, Catherine Griffin, Matthew Jackson, Christine Mézard, Gil McVean, Gregory P. Copenhaver, and Ian R Henderson (2016). "Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes". In: *PLoS Genetics* 12.7.

CLCbio (2018). *CLCbio*. URL: https://www.qiagenbioinformatics.com/ (visited on 10/01/2018).

Collier, Sarah M., Louis-Philippe Hamel, and Peter Moffett (2011). "Cell Death Mediated by the N-Terminal Domains of a Unique and Highly Conserved Class of NB-LRR Protein". In: *Molecular Plant-Microbe Interactions* 24.8. arXiv: arXiv:1408.1149. URL: http://apsjournals.apsnet.org/doi/10.1094/MPMI-03-11-0050.

Dangl, Jeffery L. and Jonathan D G Jones (June 2001). "Plant pathogens and integrated defence responses to infection." In: *Nature* 411.6839. URL: http://www.ncbi.nlm.nih.gov/pubmed/11459065.

Deslandes, L., J. Olivier, N. Peeters, D. X. Feng, M. Khounlotham, C. Boucher, I. Somssich, S. Genin, and Y. Marco (2003). "Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus". In: *Proceedings of the National Academy of Sciences* 100.13. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1230660100.

Ding, Wei, Franz Baumdicker, and Richard A. Neher (2017). "panX: pan-genome analysis and exploration". In: *Nucleic Acids Research* 46. URL: http://academic.oup.com/

nar/article/doi/10.1093/nar/gkx977/4564799/panX-pangenome-analysis-and-exploration.

Dow, Max, Marianne Newman, and Edda Von Roepenack (2000). "The Induction and Modulation of Plant Defense Responses by Bacterial Lipopolysaccharides." In: *Annual Review of Phytopathology*.

Duan, Naibin, Yang Bai, Honghe Sun, Nan Wang, Yumin Ma, Mingjun Li, Xin Wang, Chen Jiao, Noah Legall, Linyong Mao, Sibao Wan, Kun Wang, Tianming He, Shouqian Feng, Zongying Zhang, Zhiquan Mao, Xiang Shen, Xiaoliu Chen, Yuanmao Jiang, Shujing Wu, Chengmiao Yin, Shunfeng Ge, Long Yang, Shenghui Jiang, Haifeng Xu, Jingxuan Liu, Deyun Wang, Changzhi Qu, Yicheng Wang, Weifang Zuo, Li Xiang, Chang Liu, Daoyuan Zhang, Yuan Gao, Yimin Xu, Kenong Xu, Thomas Chao, Gennaro Fazio, Huairui Shu, Gan Yuan Zhong, Lailiang Cheng, Zhangjun Fei, and Xuesen Chen (2017). "Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement". In: *Nature Communications* 8.1. URL: http://dx.doi.org/10.1038/s41467-017-00336-7.

Eddy, Sean R (2011). "Accelerated profile HMM searches". In: *PLoS Computational Biology* 7.10. arXiv: NIHMS150003.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner (2009). "Single Polymerase Molecules". In: *Exchange Organizational Behavior Teaching Journal* January.

Eilbeck, Karen, Barry Moore, Carson Holt, and Mark Yandell (2009). "Quantitative measures for the management and comparison of annotated genomes". In: *BMC Bioinformatics* 10.

Elsik, Christine G., Kim C. Worley, Lan Zhang, Natalia V. Milshina, Huaiyang Jiang, Justin T. Reese, Kevin L. Childs, Anand Venkatraman, C. Michael Dickens, George M. Weinstock, and Richard A. Gibbs (2006). "Community annotation: Procedures, protocols, and supporting tools". In: *Genome Research* 16.11.

English, Adam C., Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, Donna M. Muzny, Jeffrey G. Reid, Kim C. Worley, and Richard A. Gibbs (2012). "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology". In: *PLoS ONE* 7.11.

Erbs, Gitte, Alba Silipo, Shazia Aslam, Cristina De Castro, Valeria Liparoti, Angela Flagiello, Pietro Pucci, Rosa Lanzetta, Michelangelo Parrilli, Antonio Molinaro, Mari Anne Newman, and Richard M. Cooper (2008). "Peptidoglycan and Muropeptides from Pathogens Agrobacterium and Xanthomonas Elicit Plant Innate Immunity: Structure and Activity". In: *Chemistry and Biology* 15.5.

Fauth, Markus, Patrick Schweizer, Antony Buchala, Claus Markstädter, Markus Riederer, Tadahiro Kato, and Heinrich Kauss (1998). "Cutin monomers and surface wax constituents elicit H2O2 in conditioned cucumber hypocotyl segments and enhance the activity of other H2O2 elicitors". In: *Plant Physiology* 117.

Felix, Georg, Juliana D. Duran, Sigrid Volko, and Thomas Boller (1999). "Plants have a sensitive perception system for the most conserved domain of bacterial flagellin". In: *Plant Journal* 18.3.

Felix, Georg, Martin Regenass, and Thomas Boller (1993). "Specific perception of subnanomolar concentrations of chitin fragments by tomato cells: Induction of extracellular alkalinization, changes in protein phosphorylation, and establishment of a refractory state". In: *Plant Journal* 4.2.

Ferrari, Simone (2013). "Oligogalacturonides: plant damage-associated molecular patterns and regulators of growth and development". In: *Frontiers in Plant Science* 4. URL: http://journal.frontiersin.org/article/10.3389/fpls.2013.00049/abstract.

Gao, Yuxia, Wenqiang Wang, Tian Zhang, Zhen Gong, Huayao Zhao, and Guan-Zhu Han (2018). "Out of Water: The Origin and Early Diversification of Plant R-Genes". In: *Plant Physiology* 31701091. URL: http://www.plantphysiol.org/lookup/doi/10.1104/pp.18.00185.

Gasc, Cyrielle, Eric Peyretaillade, and Pierre Peyret (Apr. 2016). "Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms." In: *Nucleic acids research*. URL: http://nar.oxfordjournals.org/content/early/2016/04/21/nar.gkw309.full.

Giolai, Michael, Pirita Paajanen, Walter Verweij, Lawrence Percival-Alwyn, David Baker, Kamil Witek, Florian Jupe, Glenn Bryan, Ingo Hein, Jonathan D G Jones, D Clark, and Matthew D Clark (2016). "Targeted capture and sequencing of gene sized DNA molecules". In: *BioTechniques* Vol. 61.No. 6.

Giolai, Michael, Pirita Paajanen, Walter Verweij, Kamil Witek, Jonathan D.G. Jones, and Matthew D. Clark (2017). "Comparative analysis of targeted long read sequencing approaches for characterization of a plant's immune receptor repertoire". In: *BMC Genomics* 18.1.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data". In: *Proceedings of the National Academy of Sciences* 108.4. arXiv: 1408.1149. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1017351108.

Golicz, Agnieszka A., Philipp E. Bayer, Guy C. Barker, Patrick P. Edger, HyeRan Kim, Paula A. Martinez, Chon Kit Kenneth Chan, Anita Severn-Ellis, W. Richard McCombie, Isobel A. P. Parkin, Andrew H. Paterson, J. Chris Pires, Andrew G. Sharpe, Haibao Tang, Graham R. Teakle, Christopher D. Town, Jacqueline Batley, David Edwards, S. Liu, I. Parkin, M. Morgante, X. Gan, J. Cao, A. A. Golicz, J. Batley, D. Edwards, W. Yao, C. N. Hirsch, Y.-H. Li, H. Tettelin, S. J. Bush, M. Schatz, R. E. Mills, B. Weckselblatt, M. K. Rudd, J. Zhang, T. Zuo, T. Peterson, K. Song, T. C.

Osborn, P. H. Williams, X. Xu, L. K. McHale, M. A. Lysak, M. A. Lysak, M. A. Koch, A. Pecinka, I. Schubert, B. Chalhoub, B. C. Meyers, A. Kozik, A. Griego, H. Kuang, R. W. Michelmore, K. Lin, T. C. Osborn, M. Tadege, M. E. Schranz, K. Okazaki, J. Zhao, S.-Y Kim, D. Xiao, S. Ridge, P. H. Brown, V. Hecht, R. G. Driessen, J. L. Weller, M. M. Kushad, D. J. Kliebenstein, V. M. Lambrix, M. Reichelt, J. Gershenzon, T. Mitchell-Olds, J. A. Hofberger, E. Lyons, P. P. Edger, J. C. Pires, M. E. Schranz, P. P. Edger, J. Zhang, G. Li, C. F. Quiros, N. M. Springer, R. A. Swanson-Wagner, K. Schneeberger, B. Langmead, S. L. Salzberg, A. V. Zimin, A. M. Bolger, M. Lohse, B. Usadel, C. Camacho, T. Arias, M. A. Beilstein, M. Tang, M. R. McKain, J. C. Pires, C. Holt, M. Yandell, I. Korf, M. Stanke, S. R. Eddy, J. Piriyapongsa, M. T. Rutledge, S. Patel, M. Borodovsky, I. K. Jordan, A. V. McDonnell, T. Jiang, A. E. Keating, B. Berger, E. B. Holub, E. Richly, J. Kurth, D. Leister, K. Howe, A. Bateman, R. Durbin, H. Li, A. Golicz, L. Li, C. J. Stoeckert, D. S. Roos, H. Tettelin, D. Riley, C. Cattuto, D. Medini, A. Rimmer, V. Obenchain, A. Stamatakis, A. Conesa, A. Alexa, J. Rahnenführer, T. Lengauer, X. Wang, K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, T. Sotelo, P. Soengas, P. Velasco, V. M. Rodríguez, and M. E. Cartea (Nov. 2016). "The pangenome of an agronomically important crop plant Brassica oleracea". In: *Nature Communications* 7. URL: http://www.nature.com/doifinder/10.1038/ncomms13390.

Gómez-Gómez, Lourdes and Thomas Boller (2002). "Flagellin perception: A paradigm for innate immunity". In: *Trends in Plant Science* 7.6.

Goritschnig, Sandra, Adam D. Steinbrenner, Derrick J. Grunwald, and Brian J. Staskawicz (2016). "Structurally distinct Arabidopsis thaliana NLR immune receptors recognize tandem WY domains of an oomycete effector". In: *New Phytologist* 210.3.

Gou, Mingyue and Jian Hua (2012). "Complex regulation of an R gene SNC1 revealed by autoimmune mutants". In: *Plant Signaling & Behavior* 7.2. URL: http://www.tandfonline.com/doi/abs/10.4161/psb.18884.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. Birren, Nir Friedman, and Aviv Regev (2013). "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data". In: *Nature Biotechnology* 29.7. arXiv: 1512.00567.

Grant, Murray R, John M McDowell, Andrew G. Sharpe, Marta De Torres Zabala, Derek J Lydiate, and Jeffery L. Dangl (1998). "Independent deletions of a pathogen-resistance gene in Brassica and Arabidopsis". In: *Proc Natl Acad Sci USA* 95.December.

Guo, Ya-Long, Joffrey Fitz, Korbinian Schneeberger, Stephan Ossowski, Jun Cao, and Detlef Weigel (Oct. 2011). "Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis." In: *Plant physiology* 157.2. URL: http://www.plantphysiol.org/content/157/2/757%20http://www.plantphysiol.org/content/157/2/757/suppl/DC1.

Gupta, Deepak K., Martin Rühl, Bagdevi Mishra, Vanessa Kleofas, Martin Hofrichter, Robert Herzog, Marek J. Pecyna, Rahul Sharma, Harald Kellner, Florian Hennicke, and Marco Thines (2018). "The genome sequence of the commercially cultivated mush-

room Agrocybe aegerita reveals a conserved repertoire of fruiting-related genes and a versatile suite of biopolymer-degrading enzymes". In: *BMC Genomics* 19.1.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler (Apr. 2013). "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt086.

Gust, Andrea A., Rory Pruitt, and Thorsten Nürnberger (2017). "Sensing Danger: Key to Activating Plant Immunity". In: *Trends in Plant Science* 22.9.

Haas, Brian J., Steven L. Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E. Allen, Joshua Orvis, Owen White, C. Robin Robin, and Jennifer R. Wortman (2008). "Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments". In: *Genome Biology* 9.

Hackl, Thomas, Rainer Hedrich, Jörg Schultz, and Frank Förster (2014). "proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus". In: *Bioinformatics* 30.21.

Hartl, D. L. (2000). "Fly meets shotgun: Shotgun wins". In: *Nature Genetics* 24.4.

Henk, Adam D., Randall F. Warren, and Roger W. Innes (1999). "A new Ac-like transposon of arabidopsis is associated with a deletion of the RPS5 disease resistance gene". In: *Genetics* 151.4.

Hofberger, Johannes A, Beifei Zhou, Haibao Tang, Jonathan D G Jones, and M Eric Schranz (Nov. 2014). "A novel approach for multi-domain and multi-gene family identification provides insights into evolutionary dynamics of disease resistance genes in core eudicot plants." In: *BMC genomics* 15.1. URL: http://www.biomedcentral.com/1471-2164/15/966.

Huddleston, John, Mark Jp Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David S Gordon, Tina A Graves-Lindsay, Katherine M Munson, Zev N Kronenberg, Laura Vives, Paul Peluso, Matthew Boitano, Chen-Shin Chin, Jonas Korlach, Richard K Wilson, and Evan E Eichler (2016). "Discovery and genotyping of structural variation from long-read haploid genome sequence data." In: *Genome research*. URL: http://www.ncbi.nlm.nih.gov/pubmed/27895111.

Illumina (2015). *Illumina sequencing technology*. URL: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html (visited on 09/10/2018).

Jacob, Florence, Saskia Vernaldi, and Takaki Maekawa (Jan. 2013). "Evolution and Conservation of Plant NLR Functions." In: *Frontiers in immunology* 4. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3782705%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Jia, Yulin, Sean A Mcadams, Gregory T Bryan, Howard P Hershey, and Barbara Valent (2000). "Direct interaction of resistance gene and avirulence gene products confers rice blast resistance". In: *The EMBO journal* 19.15.

Jiang, Hongshan, Rong Lei, Shou-Wei Ding, and Shuifang Zhu (Jan. 2014). "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads."

En. In: *BMC bioinformatics* 15.1. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-182.

Jiang, Rui, Jingcai Li, Zhendong Tian, Juan Du, Miles Armstrong, Katie Baker, Joanne Tze-Yin Lim, Jack H. Vossen, Huan He, Leticia Portal, Jun Zhou, Merideth Bonierbale, Ingo Hein, Hannele Lindqvist-Kreuze, and Conghua Xie (2018). "Potato late blight field resistance from QTL dPI09c is conferred by the NB-LRR gene R8". In: *Journal of Experimental Botany* 69.7.

Jiao, Yinping, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C. Stitzer, Bo Wang, Michael S. Campbell, Joshua C. Stein, Xuehong Wei, Chen Shan Chin, Katherine Guill, Michael Regulski, Sunita Kumari, Andrew Olson, Jonathan Gent, Kevin L. Schneider, Thomas K. Wolfgruber, Michael R. May, Nathan M. Springer, Eric Antoniou, W. Richard McCombie, Gernot G. Presting, Michael McMullen, Jeffrey Ross-Ibarra, R. Kelly Dawe, Alex Hastie, David R. Rank, and Doreen Ware (2017). "Improved maize reference genome with single-molecule technologies". In: *Nature* 546.7659. URL: http://dx.doi.org/10.1038/nature22971.

Jones, Jonathan D G and Jeffery L. Dangl (Nov. 2006). "The plant immune system." In: *Nature* 444.7117. URL: http://dx.doi.org/10.1038/nature05286.

Joubert, Aymeric, Nelly Bataille-Simoneau, Claire Campion, Thomas Guillemette, Piétrick Hudhomme, Béatrice Iacomi-Vasilescu, Thibault Leroy, Stéphanie Pochon, Pascal Poupard, and Philippe Simoneau (2011). "Cell wall integrity and high osmolarity glycerol pathways are required for adaptation of Alternaria brassicicola to cell wall stress caused by brassicaceous indolic phytoalexins". In: *Cellular Microbiology* 13.1.

Jupe, Florian, Xinwei Chen, Walter Verweij, Kamil Witek, Jonathan D G Jones, and Ingo Hein (2014). "Genomic DNA library preparation for resistance gene enrichment and sequencing (RenSeq) in plants". In: *Methods in molecular biology (Clifton, N.J.)* 1127.Dm. Ed. by Paul Birch, John T. Jones, and Jorunn I.B. Bos. URL: http://link.springer.com/10.1007/978-1-62703-986-4.

Jupe, Florian, Leighton Pritchard, Graham J Etherington, Katrin Mackenzie, Peter J A Cock, Frank Wright, Sanjeev Kumar Sharma, Dan Bolser, Glenn J Bryan, Jonathan D G Jones, and Ingo Hein (Jan. 2012). "Identification and localisation of the NB-LRR gene family within the potato genome." In: *BMC genomics* 13.1. URL: http://www.biomedcentral.com/1471-2164/13/75.

Jupe, Florian, Kamil Witek, Walter Verweij, Jadwiga Sliwka, Leighton Pritchard, Graham J Etherington, Dan Maclean, Peter J Cock, Richard M Leggett, Glenn J Bryan, Linda Cardle, Ingo Hein, and Jonathan D G Jones (Nov. 2013). "Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations." In: *The Plant journal : for cell and molecular biology* 76.3. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3935411%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Kawakatsu, Taiji, Shao-shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, Joseph R Nery, Cesar Barragan, Yupeng He, Huaming Chen, Manu Dubin, Cheng Ruei Lee, Congmao Wang, Felix Bemm, Claude Becker, Ryan O'Neil, Ronan C O'Malley, Danjuma X Quarless, Carlos Alonso-Blanco, Jorge

Andrade, Felix Bemm, Joy Bergelson, Karsten Borgwardt, Eunyoung Chae, Todd Dezwaan, Wei Ding, Joseph R Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G Grimm, Angela Hancock, Stefan R Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Chen Ruei Lee, Dazhe Meng, Todd P Michael, Richard Mott, Ni Wayan W. Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Novikova, F. Xavier Pico, Alexander Platzer, Fernando A Rabanal, Alex Rodriguez, Beth A Rowan, Patrice A. Salome, Karl Schmid, Robert J Schmitz, Ü Seren, Felice Gianluca G. Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, Xuefeng Zhou, Nicholas J. Schork, Detlef Weigel, Magnus Nordborg, and Joseph R. Ecker (2016). "Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions". In: *Cell* 166.2.

Kim, Jungeun, Chan Ju Lim, Bong Woo Lee, Jae Pil Choi, Sang Keun Oh, Raza Ahmad, Suk Yoon Kwon, Jisook Ahn, and Cheol Goo Hur (2012). "A genome-wide comparison of NB-LRR type of resistance gene analogs (RGA) in the plant Kingdom". In: *Molecules and Cells* 33.4.

Kim, Min Gab, Luis Da Cunha, Aidan J. McFall, Youssef Belkhadir, Sruti DebRoy, Jeffery L. Dangl, and David Mackey (2005). "Two Pseudomonas syringae type III effectors inhibit RIN4-regulated basal defense in Arabidopsis". In: *Cell* 121.5.

Kim, Seungill, Jieun Park, Seon-in Yeom, Yong-min Kim, Eunyoung Seo, and Ki-tae Kim (2017). "Multiple reference genome sequences of hot pepper reveal the massive evolution of plant disease resistance genes by retroduplication". In: *bioarxiv*.

Klarzynski, Olivier, Bertrand Plesse, Jean-Marie Joubert, Jean-Claude Yvin, Marguerite Kopp, Bernard Kloareg, and Bernard Fritig (2000). "Linear β-1,3 Glucans Are Elicitors of Defense Responses in Tobacco". In: *Plant Physiology* 124.3. URL: `http://www.plantphysiol.org/lookup/doi/10.1104/pp.124.3.1027`.

Koren, Sergey, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy (2017). "Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". In: *Genome research*. arXiv: `071282`.

Korf, Ian (2004). "Gene finding in novel genomes". In: *BMC Bioinformatics* 5.59.

Korlach, Jonas (2013). *Understanding Accuracy in SMRT Sequencing*. URL: `https://www.pacb.com/wp-content/uploads/2015/09/Perspective%7B%5C_%7DUnderstandingAccur` `pdf` (visited on 09/10/2018).

Kourelis, Jiorgos and Renier A. L. van der Hoorn (2018). "Defended to the Nines: 25 years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function". In: *The Plant Cell* 30.February. URL: `http://www.plantcell.org/lookup/doi/10.1105/tpc.17.00579`.

Krasileva, Ksenia V., D. Dahlbeck, and B. J. Staskawicz (2010). "Activation of an Arabidopsis Resistance Protein Is Specified by the in Planta Association of Its Leucine-Rich Repeat Domain with the Cognate Oomycete Effector". In: *the Plant Cell Online* 22.7. URL: `http://www.plantcell.org/cgi/doi/10.1105/tpc.110.075358`.

Kroj, Thomas, Emilie Chanclud, Corinne Michel-Romiti, Xavier Grand, and Jean Benoit Morel (2016). "Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread". In: *New Phytologist* 210.2.

Kuang, Hanhui, Sung-Sick Woo, Blake C. Meyers, Eviatar Nevo, and Richard W. Michelmorea (2004). "Multiple Genetic Processes Result in Heterogeneous Rates of Evolution within the Major Cluster Disease Resistance Genes in Lettuce". In: *the Plant Cell Online* 16.11. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.104.025502.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges (2017). "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times". In: *Molecular biology and evolution* 34.7.

Kunze, G. (2004). "The N Terminus of Bacterial Elongation Factor Tu Elicits Innate Immunity in Arabidopsis Plants". In: *the Plant Cell Online* 16.12. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.104.026765.

Le Roux, Clémentine, Gaëlle Huet, Alain Jauneau, Laurent Camborde, Dominique Trémousaygue, Alexandra Kraut, Binbin Zhou, Marie Levaillant, Hiroaki Adachi, Hirofumi Yoshioka, Sylvain Raffaele, Richard Berthomé, Yohann Couté, Jane E. Parker, and Laurent Deslandes (2015). "A Receptor Pair with an Integrated Decoy Converts Pathogen Disabling of Transcription Factors to Immunity". In: *Cell* 161.5. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867415004420.

Lee, DongHyuk, Gildas Bourdais, Gang Yu, Silke Robatzek, and Gitta Coaker (2015). "Phosphorylation of the Plant Immune Regulator RPM1-INTERACTING PROTEIN4 Enhances Plant Plasma Membrane H+ -ATPase Activity and Inhibits Flagellin-Triggered Immune Responses in Arabidopsis". In: *The Plant Cell* 27.7. URL: http://www.plantcell.org/lookup/doi/10.1105/tpc.114.132308.

Lee, Eduardo, Gregg a Helt, Justin T Reese, Monica C Munoz-Torres, Chris P Childers, Robert M Buels, Lincoln Stein, Ian H Holmes, Christine G Elsik, and Suzanna E Lewis (2013). "Web Apollo: a web-based genomic annotation editing platform." In: *Genome biology* 14.8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-12-research0082%20http://www.ncbi.nlm.nih.gov/pubmed/24000942.

Leibman-Markus, Meirav, Lorena Pizarro, Silvia Schuster, Z.J. Daniel Lin, Ofir Gershony, Maya Bar, Gitta Coaker, and Adi Avni (2018). "The intracellular nucleotide binding leucine-rich repeat receptor - SlNRC4a enhances immune signaling elicited by extracellular perception". In: *Plant, Cell & Environment*. URL: http://doi.wiley.com/10.1111/pce.13347.

Leister, Dario (Mar. 2004). "Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes". In: *Trends in Genetics* 20.3. URL: http://www.sciencedirect.com/science/article/pii/S0168952504000228.

Letunic, Ivica and Peer Bork (2016). "Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees". In: *Nucleic acids research* 44.W1.

Lewis, Jennifer D., Ronald Wu, David S. Guttman, and Darrell Desveaux (2010). "Allele-specific virulence attenuation of the Pseudomonas syringae HopZ1a type III effector via the Arabidopsis ZAR1 resistance protein". In: *PLoS Genetics* 6.4.

Li, Heng (n.d.). *htsbox*. URL: `https://github.com/lh3/htsbox`.

– (2018a). *seqtk Toolkit for processing sequences in FASTA/Q formats*. URL: `https://github.com/lh3/seqtk` (visited on 10/15/2018).

– (2018b). "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics*. arXiv: `1708.01492`. URL: `http://arxiv.org/abs/1708.01492`.

Li, Heng and Richard Durbin (2009a). "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14. arXiv: `1303.3997`.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin (2009b). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16. arXiv: `1006.1266v2`.

Li, Jing, Jing Ding, Wen Zhang, Yuanli Zhang, Ping Tang, Jian Qun Chen, Dacheng Tian, and Sihai Yang (2010a). "Unique evolutionary pattern of numbers of gramineous NBS-LRR genes". In: *Molecular Genetics and Genomics* 283.5.

Li, Meng, Xiqing Ma, Yi Hsuan Chiang, Koste A. Yadeta, Pengfei Ding, Liansai Dong, Yan Zhao, Xiuming Li, Yufei Yu, Ling Zhang, Qian Hua Shen, Bin Xia, Gitta Coaker, Dong Liu, and Jian Min Zhou (2014a). "Proline isomerization of the immune receptor-interacting protein RIN4 by a cyclophilin inhibits effector-triggered immunity in arabidopsis". In: *Cell Host and Microbe* 16.4. URL: `http://dx.doi.org/10.1016/j.chom.2014.09.007`.

Li, Ying-hui, Guangyu Zhou, Jianxin Ma, Wenkai Jiang, Long-guo Jin, Zhouhao Zhang, Yong Guo, Jinbo Zhang, Yi Sui, Liangtao Zheng, Shan-shan Zhang, Qiyang Zuo, Xue-hui Shi, Yan-fei Li, Wan-ke Zhang, Yiyao Hu, Guanyi Kong, Hui-long Hong, Bing Tan, Jian Song, Zhang-xiong Liu, Yaoshen Wang, Hang Ruan, Carol K L Yeung, Jian Liu, Hailong Wang, Li-juan Zhang, Rong-xia Guan, Ke-jing Wang, Wen-bin Li, Shou-yi Chen, Ru-zhen Chang, Zhi Jiang, Scott a Jackson, Ruiqiang Li, and Li-juan Qiu (2014b). "De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits". In: *Nature Biotechnology* 32.10. URL: `http://dx.doi.org/10.1038/nbt.2979`.

Li, Yingzhong, Shuxin Li, Dongling Bi, Yu Ti Cheng, Xin Li, and Yuelin Zhang (2010b). "SRFR1 negatively regulates plant NB-LRR resistance protein accumulation to prevent autoimmunity". In: *PLoS Pathogens* 6.9.

Liu, Xiaokun, Heini M. Grabherr, Roland Willmann, Dagmar Kolb, Frédéric Brunner, Ute Bertsche, Daniel Kühner, Mirita Franz-Wachtel, Bushra Amin, Georg Felix, Marc Ongena, Thorsten Nürnberger, and Andrea A. Gust (2014). "Host-induced bacterial cell wall decomposition mediates pattern-triggered immunity in Arabidopsis". In: *eLife*.

Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky (2005). "Gene identification in novel eukaryotic genomes by self-training algorithm". In: *Nucleic Acids Research* 33.20.

Loon, L.C. van, M. Rep, and C.M.J. Pieterse (2006). "Significance of Inducible Defense-related Proteins in Infected Plants". In: *Annual Review of Phytopathology* 44.1. URL: http://www.annualreviews.org/doi/10.1146/annurev.phyto.44.070505.143425.

Lozano-Durán, Rosa, Gildas Bourdais, Sheng Yang He, and Silke Robatzek (2014). "The bacterial effector HopM1 suppresses PAMP-triggered oxidative burst and stomatal immunity". In: *New Phytologist* 202.1.

Lu, H. (2003). "ACD6, a Novel Ankyrin Protein, Is a Regulator and an Effector of Salicylic Acid Signaling in the Arabidopsis Defense Response". In: *the Plant Cell Online* 15.10. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.015412.

Lupas, Andrei, M van Dyke, and Jeff Stock (1991). "Predicting coiled coils from proteins sequences". In: *Science* 252.

Mackey, David, Ben F. Holt, Aaron Wiig, and Jeffery L. Dangl (2002). "RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis". In: *Cell*.

Maekawa, Takaki, Wei Cheng, Laurentiu N. Spiridon, Armin Töller, Ewa Lukasik, Yusuke Saijo, Peiyuan Liu, Qian Hua Shen, Marius A. Micluta, Imre E. Somssich, Frank L W Takken, Andrei Jose Petrescu, Jijie Chai, and Paul Schulze-Lefert (2011). "Coiled-coil domain-dependent homodimerization of intracellular barley immune receptors defines a minimal functional module for triggering cell death". In: *Cell Host and Microbe* 9.3.

Maqbool, A., H. Saitoh, M. Franceschetti, C. E.M. Stevenson, A. Uemura, H. Kanzaki, S. Kamoun, R. Terauchi, and M. J. Banfield (2015). "Structural basis of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor". In: *eLife*.

McDonnell, A. V., T. Jiang, A. E. Keating, and Bab Berger (2006). "Paircoil2: Improved prediction of coiled coils from sequence". In: *Bioinformatics* 22.3.

McDowell, John M and Stacey A Simon (Sept. 2006). "Recent insights into R gene evolution." In: *Molecular plant pathology* 7.5. URL: http://www.ncbi.nlm.nih.gov/pubmed/20507459.

Meindl, T, T Boller, and G Felix (2000). "The bacterial elicitor flagellin activates its receptor in tomato cells according to the address-message concept." In: *Plant Cell* 12.9. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=149085%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Melotto, Maeli, William Underwood, and Sheng Yang He (2008). "Role of Stomata in Plant Innate Immunity and Foliar Bacterial Diseases". In: *Annual Review of Phytopathology* 46.

Meng, Dong, Chunlong Li, Hee-Jin Park, Jonathan Gonzalez, Jingying Wang, Abhaya M Dandekar, B Gillian Turgeon, and Lailiang Cheng (2018). "Sorbitol Modulates Resistance to Alternaria alternata by Regulating the Expression of an NLR Resistance Gene in apple". In: *The Plant cell* 30.

Meyers, Blake C, Alexander Kozik, Alyssa Griego, Hanhui Kuang, and Richard W Michelmore (2003). "Genome-wide analysis of NBS-LRR-encoding genes in Arabidop-

sis". In: *The Plant cell* 15.4. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=152331%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract%7B%5C%%7D5Cn%7B%5C%%7D3CGo%20to%20ISI%7B%5C%%7D3E://BIOSIS:PREV200300217927%7B%5C%%7D5Cnhttp://www.plantcell.org/content/15/4/809.long.

Mishra, Bagdevi, Deepak K. Gupta, Markus Pfenninger, Thomas Hickler, Ewald Langer, Bora Nam, Juraj Paule, Rahul Sharma, Bartosz Ulaszewski, Joanna Warmbier, Jaroslaw Burczyk, and Marco Thines (2018). "A reference genome of the European beech (Fagus sylvatica L.)" In: *GigaScience* 7.6.

Mondragón-Palomino, Mariana, Blake C Meyers, Richard W Michelmore, and Brandon S Gaut (Sept. 2002). "Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana." In: *Genome research* 12.9. URL: http://genome.cshlp.org/content/12/9/1305.short.

Munoz-Torres, Monica C., Justin T. Reese, Christopher P. Childers, Anna K. Bennett, Jaideep P. Sundaram, Kevin L. Childs, Juan M. Anzola, Natalia Milshina, and Christine G. Elsik (2011). "Hymenoptera Genome Database: Integrated community resources for insect species of the order Hymenoptera". In: *Nucleic Acids Research* 39.SUPPL. 1.

Narusaka, Mari, Ken Shirasu, Yoshiteru Noutoshi, Yasuyuki Kubo, Tomonori Shiraishi, Masaki Iwabuchi, and Yoshihiro Narusaka (2009). "RRS1 and RPS4 provide a dual Resistance-gene system against fungal and bacterial pathogens". In: *Plant Journal* 60.2.

Nishimura, Marc T, Ryan G Anderson, Karen A Cherkis, Terry F Law, Qingli L Liu, and Mischa Machius (2017). "TIR-only protein RBA1 recognizes a pathogen effector to regulate cell death in Arabidopsis". In: *Proceedings of the National Academy of Sciences* 114.10. URL: http://www.pnas.org/content/114/10/E2053.abstract%20http://www.pnas.org/content/114/10/E2053.full.pdf.

Noel, L. (1999). "Pronounced Intraspecific Haplotype Divergence at the RPP5 Complex Disease Resistance Locus of Arabidopsis". In: *the Plant Cell Online* 11.11. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.11.11.2099.

Ntoukakis, Vardis, Alexi L. Balmuth, Tatiana S. Mucyn, Jose R. Gutierrez, Alexandra M E Jones, and John P. Rathjen (2013). "The Tomato Prf Complex Is a Molecular Trap for Bacterial Effectors Based on Pto Transphosphorylation". In: *PLoS Pathogens* 9.1.

O'Brien, Jose A., Arsalan Daudi, Vernon S. Butt, and G. Paul Bolwell (2012). "Reactive oxygen species and their role in plant defence and cell wall metabolism". In: *Planta* 236.3.

Oerke, E-C (2006). "Crop losses to pests". In: *Journal of Agricultural Science* 144.

Ortiz, Diana and Peter N Dodds (2018). "Plant NLR Origins Traced Back to Green Algae". In: *Trends in plant science* 23.8. URL: https://doi.org/10.1016/j.tplants.2018.05.009.

Ortiz, Diana, Karine de Guillen, Stella Cesari, Véronique Chalvon, Jérome Gracy, André Padilla, and Thomas Kroj (2017). "Recognition of the Magnaporthe oryzae Effector AVR-Pia by the Decoy Domain of the Rice NLR Immune Receptor RGA5". In: *The*

Plant Cell 29.1. URL: http://www.plantcell.org/lookup/doi/10.1105/tpc.16.00435.

Peart, Jack R., Pere Mestre, Rui Lu, Isabelle Malcuit, and David C. Baulcombe (2005). "NRG1, a CC-NB-LRR protein, together with N, a TIR-NB-LRR protein, mediates resistance against tobacco mosaic virus". In: *Current Biology* 15.10.

Peele, Hanneke M, Na Guan, Johan Fogelqvist, and Christina Dixelius (2014). "Loss and retention of resistance genes in five species of the Brassicaceae family". In: *BMC Plant Biology* 14. URL: http://www.biomedcentral.com/1471-2229/14/298.

Phillippy, Adam, Sergey Koren, and Brian Walenz (2018). *Canu Parameter Reference*. URL: https://canu.readthedocs.io/en/latest/parameter-reference.html%7B%5C#%7Dparameter-reference (visited on 11/19/2018).

Plomion, Christophe, Jean-Marc Aury, Joëlle Amselem, Thibault Leroy, Florent Murat, Sébastien Duplessis, Sébastien Faye, Nicolas Francillonne, Karine Labadie, Grégoire Le Provost, Isabelle Lesur, Jérôme Bartholomé, Patricia Faivre-Rampant, Annegret Kohler, Jean-Charles Leplé, Nathalie Chantret, Jun Chen, Anne Diévart, Tina Alaeitabar, Valérie Barbe, Caroline Belser, Hélène Bergès, Catherine Bodénès, Marie-Béatrice Bogeat-Triboulot, Marie-Lara Bouffaud, Benjamin Brachi, Emilie Chancerel, David Cohen, Arnaud Couloux, Corinne Da Silva, Carole Dossat, François Ehrenmann, Christine Gaspin, Jacqueline Grima-Pettenati, Erwan Guichoux, Arnaud Hecker, Sylvie Herrmann, Philippe Hugueney, Irène Hummel, Christophe Klopp, Céline Lalanne, Martin Lascoux, Eric Lasserre, Arnaud Lemainque, Marie-Laure Desprez-Loustau, Isabelle Luyten, Mohammed-Amin Madoui, Sophie Mangenot, Clémence Marchal, Florian Maumus, Jonathan Mercier, Célia Michotey, Olivier Panaud, Nathalie Picault, Nicolas Rouhier, Olivier Rué, Camille Rustenholz, Franck Salin, Marçal Soler, Mika Tarkka, Amandine Velt, Amy E. Zanne, Francis Martin, Patrick Wincker, Hadi Quesneville, Antoine Kremer, and Jérôme Salse (2018). "Oak genome reveals facets of long lifespan". In: *Nature Plants* iDiv. URL: http://www.nature.com/articles/s41477-018-0172-3.

Porter, Brad W, Maya Paidi, Ray Ming, Maqsudul Alam, Wayne T Nishijima, and Yun J Zhu (2009). "Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family." In: *Molecular genetics and genomics : MGG* 281.6. URL: http://www.ncbi.nlm.nih.gov/pubmed/19263082.

Pritchard, Leighton and Paul R J Birch (Dec. 2014). "The zigzag model of plant-microbe interactions: is it time to move on?" In: *Molecular plant pathology* 15.9. URL: http://www.ncbi.nlm.nih.gov/pubmed/25382065.

Pucker, Boas, Daniela Holtgraewe, Kai Bernd Stadermann, Katharina Frey, Bruno Huettel, Richard Reinhardt, and Bernd Weisshaar (2018). "A Chromosome-level Sequence Assembly Reveals the Structure of the Arabidopsis thaliana Nd-1 Genome and its Gene Set". In: *bioRxiv*. URL: https://www.biorxiv.org/content/early/2018/09/06/407627.

Qi, D., B. J. DeYoung, and R. W. Innes (2012). "Structure-Function Analysis of the Coiled-Coil and Leucine-Rich Repeat Domains of the RPS5 Disease Resistance Protein". In: *Plant Physiology* 158.4. URL: http://www.plantphysiol.org/cgi/doi/10.1104/pp.112.194035.

Qi, D., U. Dubiella, S. H. Kim, D. I. Sloss, R. H. Dowen, J. E. Dixon, and R. W. Innes (2014). "Recognition of the Protein Kinase AVRPPHB SUSCEPTIBLE1 by the Disease Resistance Protein RESISTANCE TO PSEUDOMONAS SYRINGAE5 Is Dependent on S-Acylation and an Exposed Loop in AVRPPHB SUSCEPTIBLE1". In: *PLANT PHYSIOLOGY*.

Quail, Michael A, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers". In: *BMC genomics* 13.341. arXiv: NIHMS150003.

Qutob, D., B. Kemmerling, F. Brunner, I. Kufner, S. Engelhardt, A. A. Gust, B. Luberacki, H. U. Seitz, D. Stahl, T. Rauhut, E. Glawischnig, G. Schween, B. Lacombe, N. Watanabe, E. Lam, R. Schlichting, D. Scheel, K. Nau, G. Dodt, D. Hubert, M. Gijzen, and T. Nurnberger (2006). "Phytotoxicity and Innate Immune Responses Induced by Nep1-Like Proteins". In: *the Plant Cell Online* 18.12. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.106.044180.

Ravensdale, Michael, Maud Bernoux, Thomas Ve, Bostjan Kobe, Peter H. Thrall, Jeffrey G. Ellis, and Peter N. Dodds (2012). "Intramolecular Interaction Influences Binding of the Flax L5 and L6 Resistance Proteins to their AvrL567 Ligands". In: *PLoS Pathogens* 8.11.

Robatzek, Silke, Pascal Bittel, Delphine Chinchilla, Petra Köchner, Georg Felix, Shin Han Shiu, and Thomas Boller (2007). "Molecular identification and characterization of the tomato flagellin receptor LeFLS2, an orthologue of Arabidopsis FLS2 exhibiting characteristically different perception specificities". In: *Plant Molecular Biology* 64.5.

Rogers, Elizabeth, J Glazebrook, and Frederick Ausubel (1996). "Mode of action of the Arabidopsis thaliana phytoalexin camalexin and its role in Arabidopsis-pathogen interactions". In: *Molecular plant-microbe interactions : MPMI* 9.

Rose, Laura E, Peter D Bittner-eddy, Charles H Langley, Eric B Holub, Richard W Michelmore, and Jim L Beynon (2004). "The Maintenance of Extreme Amino Acid Diversity at the Disease Resistance Gene, RPP13, in Arabidopsis thaliana". In: *Genetics* 166.2.

Sarris, Panagiotis F., Zane Duxbury, Sung Un Huh, Yan Ma, Cécile Segonzac, Jan Sklenar, Paul Derbyshire, Volkan Cevik, Ghanasyam Rallapalli, Simon B. Saucet, Lennart Wirthmueller, Frank L.H. Menke, Kee Hoon Sohn, and Jonathan D.G. Jones (2015). "A plant immune receptor detects pathogen effectors that target WRKY transcription factors". In: *Cell* 161.5.

Sarris, Panagiotis F, Volkan Cevik, Gulay Dagdas, Jonathan D G Jones, and Ksenia V. Krasileva (2016). "Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens." In: *BMC biology* 14.1. URL: http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-016-0228-7.

Sawinski, Katja, Sophia Mersmann, Silke Robatzek, and Maik Böhmer (2013). "Guarding the Green: Pathways to Stomatal Immunity". In: *Molecular Plant-Microbe Interactions* 26.6. URL: http://apsjournals.apsnet.org/doi/10.1094/MPMI-12-12-0288-CR.

Schirmer, Melanie, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince (2016). "Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data". In: *BMC Bioinformatics* 17.1. URL: http://dx.doi.org/10.1186/s12859-016-0976-y.

Seo, Jeong-Sun, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, Han Cao, Ji-Young Yun, Jihye Kim, Junho Kuk, Gun Hwa Park, Juhyeok Kim, Hanna Ryu, Jongbum Kim, Mira Roh, Jeonghun Baek, Michael W. Hunkapiller, Jonas Korlach, Jong-Yeon Shin, and Changhoon Kim (Oct. 2016). "De novo assembly and phasing of a Korean human genome". In: *Nature* 538.7624. URL: http://www.nature.com/doifinder/10.1038/nature20098.

Seto, Derek, Noushin Koulena, Timothy Lo, Alexandra Menna, David S. Guttman, and Darrell Desveaux (2017). "Expanded type III effector recognition by the ZAR1 NLR protein using ZED1-related kinases". In: *Nature Plants* 3.March.

Shao, Zhu-Qing, Jia-Yu Xue, Ping Wu, Yan-Mei Zhang, Yue Wu, Yue-Yu Hang, Bin Wang, and Jian-Qun Chen (Apr. 2016). "Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns." In: *Plant physiology* 170.4. URL: http://www.ncbi.nlm.nih.gov/pubmed/26839128%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4825152.

Shao, Zhu-Qing, Yan-Mei Zhang, Yue-Yu Hang, Jia-Yu Xue, Guang-Can Zhou, Ping Wu, Xiao-Yi Wu, Xun-Zong Wu, Qiang Wang, Bin Wang, and Jianqun Chen (2014). "Long-Term Evolution of Nucleotide-Binding Site-Leucine-Rich Repeat (NBS-LRR) Genes: Understandings Gained From and Beyond the Legume Family." In: *Plant physiology* 166.September. URL: http://www.ncbi.nlm.nih.gov/pubmed/25052854.

Slater, Guy St C. and Ewan Birney (2005). "Automated generation of heuristics for biological sequence comparison". In: *BMC Bioinformatics* 6.

Solovyev, Victor, Peter Kosarev, Igor Seledsov, and Denis Vorobyev (2006). "Automatic annotation of eukaryotic genes, pseudogenes and promoters." In: *Genome biology* 7 Suppl 1.Suppl 1. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1810547/pdf/gb-2006-7-s1-s10.pdf.

Stam, Remco, Daniela Scheikl, and Aurelien Tellier (2016). "Pooled Enrichment Sequencing Identifies Diversity and Evolutionary Pressures at NLR Resistance Genes within a Wild Tomato Population". In: *Genome biology and evolution* 8.5.

Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern (July 2004). "AUGUSTUS: a web server for gene finding in eukaryotes." In: *Nucleic acids research* 32.Web Server issue. URL: http://nar.oxfordjournals.org/content/32/suppl%7B%5C_%7D2/W309.full.

Stein, Joshua C., Yeisoo Yu, Dario Copetti, Derrick J. Zwickl, Li Zhang, Chengjun Zhang, Kapeel Chougule, Dongying Gao, Aiko Iwata, Jose Luis Goicoechea, Sharon Wei, Jun Wang, Yi Liao, Muhua Wang, Julie Jacquemin, Claude Becker, Dave Kudrna, Jianwei Zhang, Carlos E.M. Londono, Xiang Song, Seunghee Lee, Paul Sanchez, Andrea Zuccolo, Jetty S.S. Ammiraju, Jayson Talag, Ann Danowitz, Luis F. Rivera, Andrea R. Gschwend, Christos Noutsos, Cheng Chieh Wu, Shu Min Kao, Jhih Wun Zeng, Fu Jin Wei, Qiang Zhao, Qi Feng, Moaine El Baidouri, Marie Christine Car-

pentier, Eric Lasserre, Richard Cooke, Daniel Da Rosa Farias, Luciano Carlos Da Maia, Railson S. Dos Santos, Kevin G. Nyberg, Kenneth L. McNally, Ramil Mauleon, Nickolai Alexandrov, Jeremy Schmutz, Dave Flowers, Chuanzhu Fan, Detlef Weigel, Kshirod K. Jena, Thomas Wicker, Mingsheng Chen, Bin Han, Robert Henry, Yue Ie C. Hsing, Nori Kurata, Antonio Costa De Oliveira, Olivier Panaud, Scott A. Jackson, Carlos A. Machado, Michael J. Sanderson, Manyuan Long, Doreen Ware, and Rod A. Wing (2018). "Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza". In: *Nature Genetics* 50.2. URL: http://dx.doi.org/10.1038/s41588-018-0040-0.

Steinbrenner, Adam D., Sandra Goritschnig, and Brian J. Staskawicz (2015). "Recognition and Activation Domains Contribute to Allele-Specific Responses of an Arabidopsis NLR Receptor to an Oomycete Effector Protein". In: *PLOS Pathogens* 11.2. URL: http://dx.plos.org/10.1371/journal.ppat.1004665.

Steuernagel, Burkhard, Florian Jupe, Kamil Witek, Jonathan D G Jones, and Brande B H Wulff (Jan. 2015). "NLR-parser: Rapid annotation of plant NLR complements." In: *Bioinformatics (Oxford, England)*. URL: http://www.ncbi.nlm.nih.gov/pubmed/25586514.

Stuttmann, Johannes, Nora Peine, Ana V. Garcia, Christine Wagner, Sayan R. Choudhury, Yiming Wang, Geo Velikkakam James, Thomas Griebel, Ruben Alcázar, Kenichi Tsuda, Korbinian Schneeberger, and Jane E. Parker (2016). "Arabidopsis thaliana DM2h (R8) within the Landsberg RPP1-like Resistance Locus Underlies Three Different Cases of EDS1-Conditioned Autoimmunity". In: *PLoS Genetics* 12.4.

Teh, Bin Tean, Kevin Lim, Chern Han Yong, Cedric Chuan Young Ng, Sushma Ramesh Rao, Vikneswari Rajasegaran, Weng Khong Lim, Choon Kiat Ong, Ki Chan, Vincent Kin Yuen Cheng, Poh Sheng Soh, Sanjay Swarup, Steven G Rozen, Niranjan Nagarajan, and Patrick Tan (2017). "The draft genome of tropical fruit durian (Durio zibethinus)". In: *Nature Genetics* 49.11. URL: http://www.nature.com/doifinder/10.1038/ng.3972.

Thomma, Bart P H J, Thorsten Nürnberger, and Matthieu H A J Joosten (Jan. 2011). "Of PAMPs and effectors: the blurred PTI-ETI dichotomy." In: *The Plant cell* 23.1. URL: http://www.plantcell.org/content/23/1/4.

Tian, D., H. Araki, E. Stahl, J. Bergelson, and M. Kreitman (2002). "Signature of balancing selection in Arabidopsis". In: *Proceedings of the National Academy of Sciences* 99.17. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.172203599.

Todesco, Marco, Sureshkumar Balasubramanian, Tina T Hu, M Brian Traw, Matthew Horton, Petra Epple, Christine Kuhns, Sridevi Sureshkumar, Christopher Schwartz, Christa Lanz, Roosa A.E. Laitinen, Yu Huang, Joanne Chory, Volker Lipka, Justin O Borevitz, Jeffery L. Dangl, Joy Bergelson, Magnus Nordborg, and Detlef Weigel (2010). "Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana". In: *Nature* 465.7298. URL: http://dx.doi.org/10.1038/nature09083.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nature Protocols* 7.3.

TwoBlades Foundation (2017). *Resources*. URL: http://2blades.org/resources/ (visited on 09/26/2018).

Tyagi, M, A G de Brevern, N Srinivasan, and B Offmann (2008). "Protein structure mining using a structural alphabet". In: *Proteins* 71.2.

Ueda, Hirokazu, Yube Yamaguchi, and Hiroshi Sano (2006). "Direct interaction between the tobacco mosaic virus helicase domain and the ATP-bound resistance protein, N factor during the hypersensitive response in tobacco plants". In: *Plant Molecular Biology* 61.1-2.

Van Weymers, Pauline S. M., Katie Baker, Xinwei Chen, Brian Harrower, David E. L. Cooke, Eleanor M. Gilroy, Paul R. J. Birch, Gaëtan J. A. Thilliez, Alison K. Lees, James S. Lynott, Miles R. Armstrong, Gaynor McKenzie, Glenn J. Bryan, and Ingo Hein (2016). "Utilizing "Omic" Technologies to Identify and Prioritize Novel Sources of Resistance to the Oomycete Pathogen Phytophthora infestans in Potato Germplasm Collections". In: *Frontiers in Plant Science* 7.May. URL: http://journal.frontiersin.org/Article/10.3389/fpls.2016.00672/abstract.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, and Ashlee M. Earl (2014). "Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement". In: *PLoS ONE* 9.11.

Wan, Hongjian, Wei Yuan, Kailiang Bo, Jia Shen, Xin Pang, and Jinfeng Chen (2013). "Genome-wide analysis of NBS-encoding disease resistance genes in Cucumis sativus and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops". In: *BMC genomics* 14.

Wang, Guoxun, Brice Roux, Feng Feng, Endrick Guy, Lin Li, Nannan Li, Xiaojuan Zhang, Martine Lautier, Marie Françoise Jardinaud, Matthieu Chabannes, Matthieu Arlat, She Chen, Chaozu He, Laurent D. Noël, and Jian Min Zhou (2015). "The Decoy Substrate of a Pathogen Effector and a Pseudokinase Specify Pathogen-Induced Modified-Self Recognition and Immunity in Plants". In: *Cell Host and Microbe*.

Wang, Jie (2018). *deFusion*. URL: https://wjidea.github.io/defusion/ (visited on 10/05/2018).

Wang, K., M. Senthil-Kumar, C.-M. Ryu, L. Kang, and K. S. Mysore (2012). "Phytosterols Play a Key Role in Plant Innate Immunity against Bacterial Pathogens by Regulating Nutrient Efflux into the Apoplast". In: *Plant Physiology* 158.4. URL: http://www.plantphysiol.org/cgi/doi/10.1104/pp.111.189217.

Waterhouse, Robert M., Fredrik Tegenfeldt, Jia Li, Evgeny M. Zdobnov, and Evgenia V. Kriventseva (2013). "OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs". In: *Nucleic Acids Research* 41.D1. arXiv: 1311.4706.

Williams, Simon J., Kee Hoon Sohn, Lin Wan, Maud Bernoux, Panagiotis F. Sarris, Cecile Segonzac, Thomas Ve, Yan Ma, Simon B. Saucet, Daniel J. Ericsson, Lachlan W Casey, Thierry Lonhienne, Donald J Winzor, Xiaoxiao Zhang, Anne Coerdt, Jane E. Parker, Peter N. Dodds, Bostjan Kobe, and Jonathan D.G. Jones (2014). "Structural

Basis for Assembly and Function of a Heterodimeric Plant Immune Receptor". In: *Science* 344.April.

Williams, Simon J., Pradeep Sornaraj, Emma DeCourcy-Ireland, R. Ian Menz, Bostjan Kobe, Jeffrey G. Ellis, Peter N. Dodds, and Peter A. Anderson (2011). "An Autoactive Mutant of the M Flax Rust Resistance Protein Has a Preference for Binding ATP, Whereas Wild-Type M Protein Binds ADP". In: *Molecular Plant-Microbe Interactions* 24.8. URL: http://apsjournals.apsnet.org/doi/10.1094/MPMI-03-11-0052.

Witek, Kamil, Florian Jupe, Agnieszka I. Witek, David Baker, Matthew D. Clark, and Jonathan D.G. Jones (2016a). "Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing". In: *Nature Biotechnology* 34.6. URL: http://dx.doi.org/10.1038/nbt.3540.

Witek, Kamil, Florian Jupe, Agnieszka I Witek, David Baker, Matthew D Clark, and Jonathan D G Jones (Apr. 2016b). "Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing". In: *Nature Biotechnology* advance on. URL: http://dx.doi.org/10.1038/nbt.3540.

Wu, Chih Hang, Lida Derevnina, and Sophien Kamoun (2018). "Receptor networks underpin plant immunity". In: *Science* 360.6395.

Wu, Chih-Hang, Ksenia V. Krasileva, Mark J. Banfield, Ryohei Terauchi, and Sophien Kamoun (2015). "The "sensor domains" of plant NLR proteins: more than decoys?" In: *Frontiers in Plant Science* 6.March. URL: http://journal.frontiersin.org/Article/10.3389/fpls.2015.00134/abstract.

Xiao, Shunyuan, Ozer Calis, Elaine Patrick, Guangmin Zhang, Piyavadee Charoenwattana, Paul Muskett, Jane E. Parker, and John G. Turner (2005). "The atypical resistance gene, RPW8, recruits components of basal defence for powdery mildew resistance in Arabidopsis". In: *Plant Journal* 42.1.

Xiao, Shunyuan, Simon Ellwood, Ozer Calis, Elaine Patrick, Tianxian Li, Mark Coleman, John G Turner, Shunyuan Xiao, Simon Ellwood, Ozer Callis, Elaine Patrick, Tianxian Li, Mark Coleman, and John G Turner (2001). "Broad-Spectrum mildew resistance in Arabidopsis mediated by RPW8". In: *Science* 291.5501.

Yamaguchi, T, a Yamada, N Hong, T Ogawa, T Ishii, and N Shibuya (2000). "Differences in the recognition of glucan elicitor signals between rice and soybean: beta-glucan fragments from the rice blast disease fungus Pyricularia oryzae that elicit phytoalexin biosynthesis in suspension-cultured rice cells." In: *The Plant cell* 12.5.

Yandell, Mark and Daniel Ence (2012). "A beginner's guide to eukaryotic genome annotation". In: *Nature Reviews Genetics* 13.5. URL: http://dx.doi.org/10.1038/nrg3174.

Yang, Sihai, Zhumei Feng, Xiuyan Zhang, Ke Jiang, Xinqing Jin, Yueyu Hang, Jian-Qun Chen, and Dacheng Tian (2006). "Genome-wide investigation on the genetic variations of rice disease resistance genes". In: *Plant Molecular Biology* 62.

Yi, Hankuil and Eric J Richards (2007). "A Cluster of Disease Resistance Genes in Arabidopsis Is Coordinately Regulated by Transcriptional Activation and RNA Silencing". In: *THE PLANT CELL ONLINE* 19.9. URL: http://www.plantcell.org/cgi/doi/10.1105/tpc.107.051821.

Yu, Jingyin, Sadia Tehrim, Fengqi Zhang, Chaobo Tong, Junyan Huang, Xiaohui Cheng, Caihua Dong, Yanqiu Zhou, Rui Qin, Wei Hua, and Shengyi Liu (2014). "Genome-wide comparative analysis of NBS-encoding genes between Brassica species and Arabidopsis thaliana." In: *BMC genomics* 15.1. URL: http://www.biomedcentral.com/1471-2164/15/3.

Yue, Jia-Xing, Blake C Meyers, Jian-Qun Chen, Dacheng Tian, and Sihai Yang (Mar. 2012). "Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes." In: *The New phytologist* 193.4. URL: http://www.ncbi.nlm.nih.gov/pubmed/22212278.

Zapata, Luis, Jia Ding, Eva-Maria Willing, Benjamin Hartwig, Daniela Bezdan, Wen-Biao Jiao, Vipul Patel, Geo Velikkakam James, Maarten Koornneef, Stephan Ossowski, and Korbinian Schneeberger (July 2016). "Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.28. URL: http://www.ncbi.nlm.nih.gov/pubmed/27354520%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4948326.

Zdobnov, E M and R Apweiler (Sept. 2001). "InterProScan–an integration platform for the signature-recognition methods in InterPro". In: *Bioinformatics* 17.9.

Zhang, Jie, Wei Li, Tingting Xiang, Zixu Liu, Kristin Laluk, Xiaojun Ding, Yan Zou, Minghui Gao, Xiaojuan Zhang, She Chen, Tesfaye Mengiste, Yuelin Zhang, and Jian Min Zhou (2010a). "Receptor-like cytoplasmic kinases integrate signaling from multiple plant immune receptors and are targeted by a Pseudomonas syringae effector". In: *Cell Host and Microbe* 7.4. URL: http://dx.doi.org/10.1016/j.chom.2010.03.007.

Zhang, Jie, Feng Shao, Yan Li, Haitao Cui, Linjie Chen, Hongtao Li, Yan Zou, Chengzu Long, Lefu Lan, Jijie Chai, She Chen, Xiaoyan Tang, and Jian Min Zhou (2007). "A Pseudomonas syringae Effector Inactivates MAPKs to Suppress PAMP-Induced Immunity in Plants". In: *Cell Host and Microbe* 1.3.

Zhang, Meiping, Yen Hsuan Wu, Mi Kyung Lee, Yun Hua Liu, Ying Rong, Teofila S. Santos, Chengcang Wu, Fangming Xie, Randall L. Nelson, and Hong Bin Zhang (2010b). "Numbers of genes in the NBS and RLK families vary by more than four-fold within a plant species and are regulated by multiple factors". In: *Nucleic Acids Research* 38.19.

Zhang, Rongzhi, Florent Murat, Caroline Pont, Thierry Langin, and Jerome Salse (2014). "Paleo-evolutionary plasticity of plant disease resistance genes". In: *BMC Genomics* 15.1. URL: http://www.biomedcentral.com/1471-2164/15/187.

Zhang, Yu, Rui Xia, Hanhui Kuang, and Blake C Meyers (2016). "The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them". In: *Molecular biology and evolution* 33.10.

Zhang, Zhibin, Yaling Wu, Minghui Gao, Jie Zhang, Qing Kong, Yanan Liu, Hongping Ba, Jianmin Zhou, and Yuelin Zhang (2012). "Disruption of PAMP-induced MAP kinase cascade by a pseudomonas syringae effector activates plant immunity mediated by the NB-LRR protein SUMM2". In: *Cell Host and Microbe* 11.3. URL: http://dx.doi.org/10.1016/j.chom.2012.01.015.

Zheng, Fengya, Haiyang Wu, Rongzhi Zhang, Shiming Li, Weiming He, Fuk-Ling Wong, Genying Li, Shancen Zhao, and Hon-Ming Lam (2016). "Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family". In: *BMC Genomics* 17.1. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2736-9.

Zhong, Yan, Huan Yin, Daniel James Sargent, Mickael Malnoy, and Zong Ming Cheng (2015). "Species-specific duplications driving the recent expansion of NBS-LRR genes in five Rosaceae species". In: *BMC Genomics* 16.1.

Zhu, Wangsheng, Maricris Zaidem, Anna-Lena Van de Weyer, Rafal M Gutaker, Sangtae Kim, Felix Bemm, Lei Li, Frederik Unger, Marcel Janis, Monika Demar, and Detlef Weigel (2018). "Modulation of ACD6 dependent hyperimmunity by natural alleles of an Arabidopsis thaliana NLR resistance gene". In: *PLoS Genetics* 14.9.

Appendices

A. Supplementary Tables

Table A.1.: The table shows for each accession the used 'Identifier' and the corresponding 1001 Identifier (1001 id). The accession name (Accession) is given, as well as the seed stock numbers if known (Stock). The country (Origin) of each accession is given together with the coordinates (lat and long). It is shown if an accession belongs to the 'Relict' group or is a founder of the MAGIC-lines (MAGIC).

Identifier	1001_id	Accession	Stock	Origin	lat	long	Relict	MAGIC
5784	5784	Ty-1	CS78790	UK	56.4	-5.2	0	0
6981	6981	Ws-2	CS76631	RUS	52.3	30	0	0
9134	9134	Yeg-8	CS75475	ARM	39.87	45.36	0	0
9610	9610	Lesno-4	CS77034	RUS	53.04	51.96	0	0
10015	10015	Sha	CS22690	AFG	37.29	71.3	0	0
5993	5993	DraIV 6-22	CS76823	CZE	49.41	16.28	0	0
9669	9669	Mitterberg-2-185	CS77086	ITA	46.37	11.28	0	0
9784	9784	Erg2-6	CS76845	GER	48.5	8.8	0	0
9792	9792	Lu4-2	CS77058	GER	48.54	9.09	0	0
1925	1925	MNF-Che-2	CS76185	USA	43.53	-86.18	0	0
6909	6909	Col-0	CS22681	USA	38.3	-92.3	0	1
9100	9100	Lag1-2	CS75441	GEO	41.83	46.28	0	0
9658	9655	Marce-1	#N/A	ITA	38.92	16.47	0	0
9721	9721	Schip-1	CS77239	BUL	42.72	25.33	0	0
9533	9533	Cem-0	CS76763	ESP	41.15	-4.32	1	0
9542	9542	Fun-0	CS76872	ESP	40.79	-4.05	1	0
9550	9550	Iso-4	CS7694	ESP	43.05	-5.37	1	0
9554	9554	Lso-0	CS77055	ESP	38.86	-3.16	1	0
9600	9600	Vis-0	CS78848	ESP	39.85	-6.04	1	0
9518	9518	Alm-0	CS76660	ESP	39.88	-0.36	1	0
9537	9537	Cum-1	CS76787	ESP	38.07	-6.66	0	0
9557	9557	Moa-0	CS77102	ESP	42.46	0.7	0	0
9597	9597	Vig-1	CS78843	ESP	42.31	-2.53	0	0
6899	6899	Bay-0	CS22676	GER	49	11	0	0
6906	6906	C24	CS22680	POR	40.21	-8.43	0	0
6911	6911	Cvi-0	CS76789	CPV	15.11	-23.62	1	0
9580	9580	Scm-0	CS77241	ESP	38.68	-3.57	0	0
9654	9654	Liri-1	CS77041	ITA	41.41	13.77	0	0
108	108	LDV-18	CS77013	FRA	48.52	-4.07	0	0
9911	9928	BEZ-9	#N/A	FRA	44.12	3.77	0	0
6981.2	6981	Ws-2	CS28828	RUS	52.3	30	0	0
7058	7058	Bur-0	CS28124	IRL	54.1	-6.2	0	1
7111	7111	Edi-0	CS28220	UK	55.95	-3.16	0	1
7213	7213	Ler-0	CS28445	GER	47.98	10.87	0	1
7288	7288	Oy-0	CS28591	NOR	60.39	6.19	0	1

Table A.1.: continued

Identifier	1001_id	Accession	Stock	Origin	lat	long	Relict	MAGIC
7373	7373	Tsu-0	CS28780	JPN	34.43	136.31	0	1
7067	7067	Ct-1	CS28195	ITA	37.3	15	0	1
7186	7186	Kn-0	CS28395	LTU	54.9	23.89	0	1
7273	7273	No-0	CS28565	GER	51.06	13.3	0	1
7396	7396	Ws-0	CS28824	RUS	52.3	30	0	1
7413	7413	Wil-2	TSL-JJ-SP2486	LTU	54.68	25.32	0	1
6909.2	6909	Col-0	CS28167	USA	38.3	-92.3	0	0
7322	7322	Rsch-4	CS28716	RUS	56.3	34	0	1
7415	7415	Wu-0	N6897	GER	49.79	9.94	0	1
7416	7416	Yo-0	CS22624	USA	37.45	-119.35	0	0
7063.2	7063	Can-0	CS28130	ESP	29.21	-13.48	1	1
7328	7328	Sf-2	CS28731	ESP	41.78	3.03	0	1
6939	6939	Mt-0	N1380	LIB	32.34	22.46	0	1
7167	7167	Hi-0	CS28346	NED	52	5	0	1
7308	7308	Po-0	CS28648	GER	50.72	7.1	0	1
6924	6924	HR-5	CS22596	UK	51.41	-0.64	0	0
7417	7417	Zu-0	N6902	SUI	47.37	8.55	0	1
9536	9536	Cor-0	CS76782	ESP	40.83	-2	1	0
7063	7186	Kn-0	#N/A	LTU	54.9	23.89	0	1
9543	9543	Gra-0	CS76886	ESP	36.77	-5.39	1	0
9545	9545	Her-12	CS76920	ESP	39.4	-5.78	1	0
9549	9549	Hum-2	CS76943	ESP	42.23	-3.69	1	0
9583	9583	Sne-0	CS77258	ESP	37.09	-3.38	1	0
9837	9837	Con-0	CS76780	ESP	37.94	-5.6	1	0
9871	9871	Nac-0	CS77117	ESP	40.75	-3.99	1	0
9944	9944	Don-0	CS76411	ESP	36.83	-6.36	1	0
6974	6974	Ull2-5	CS78818	SWE	56.06	13.97	0	0
9555	9555	Mar-1	CS77068	ESP	39.58	-3.93	1	0
9598	9598	Vim-0	CS78844	ESP	41.88	-6.51	1	0
9905	9905	Ven-0	CS78840	ESP	40.76	-4.01	1	0
9762	9762	Etna-2	CS76487	ITA	37.69	14.98	1	0
9764	9764	Qar-8a	CS76581	LBN	34.1	35.84	1	0
9332	9332	Bar-1	CS76688	SWE	62.87	18.38	0	0
9869	9869	Moj-0	CS77105	ESP	36.76	-5.28	1	0
9879	9879	Per-0	CS77169	ESP	37.6	-1.12	1	0
9887	9887	Pun-0	CS77196	ESP	40.4	-4.77	1	0
9947	9947	Ped-0	CS76415	ESP	40.74	-3.9	1	0
9832	9832	Cat-0	CS76759	ESP	40.54	-3.69	1	0

Table A.2.: The table shows for each accession (Identifier),which size selection method was used (Size_Sel: BP=BluePippin, SE=SageElf). The Sequencing Provider (Seq_Prov) (MPI=Max Planck Institute for Developmental Biology, Tuebingen, EI=Earlham Institute Norwich, UNC=University of Chapel Hill) and the Sequencing Facility (Seq_Fac) are given, and the used 'Library Adaptors' are shown (for custom adaptor sequences see table 3.B.1). The table contains the number of sequenced SMRT cells (cells), and if an accession was sequenced multiplexed (multi). It also shows for which accessions PCRfree whole genome sequencing short read data (SR) was produced.

Identifier	Size_Sel	Seq_Prov	Seq_Fac	Library_Adaptors	cells	multi	SR
5784	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6981	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9134	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9610	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7351	1	No	1
10015	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
5993	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9669	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7354	1	No	1
9784	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9792	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
1925	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	2	No	1
6909	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9100	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9658	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7353	1	No	1
9721	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7355	1	No	1
9533	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9542	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9550	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9554	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9600	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9518	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9537	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9557	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9597	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6899	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6906	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6911	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	NA
9580	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
9654	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7352	1	No	1
108	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	2	No	1
9911	BP	MPI	MPI/MPGC_Cologne	NEBNext_E7350	1	No	1
6981.2	SE	EI	TSL	AJI_6+OF-PacF6		Yes	NA
7058	SE	EI	TSL	AJI_1+OF-PacF1	8	Yes	1
7111	SE	EI	TSL	AJI_8+OF-PacF8	5	Yes	1
7213	SE	EI	TSL	AJI_3+OF-PacF3	5	Yes	1
7288	SE	EI	TSL	AJI_1	1	No	1
7373	SE	EI	TSL	AJI_5+OF-PacF5	8	Yes	1
7067	SE	EI	TSL	AJI_3+OF-PacF3	8	Yes	1
7186	SE	EI	TSL	AJI_2+OF-PacF2	5	Yes	1
7273	SE	EI	TSL	AJI_4+OF-PacF4	5	Yes	1

Table A.2.: continued

Identifier	Size_Sel	Seq_Prov	Seq_Fac	Library_Adaptors	cells	multi	SR
7396	SE	EI	TSL	AJI_6+OF-PacF6	8	Yes	1
7413	SE	EI	TSL	AJI_1	1	No	1
6909.2	SE	EI	TSL	Nextera1F+1R			NA
7322	SE	EI	TSL	AJI_5+OF-PacF5	5	Yes	1
7415	SE	EI	TSL	AJI_7+OF-PacF7	5	Yes	1
7416	SE	EI	TSL	AJI_1			1
7063.2	SE	EI	TSL	AJI_2+OF-PacF2	1	Yes	NA
7328	SE	EI	TSL	AJI_1	1	No	1
6939	SE	EI	TSL	AJI_9+OF-PacF9	5	Yes	1
7167	SE	EI	TSL	AJI_1+OF-PacF1	5	Yes	1
7308	SE	EI	TSL	AJI_4+OF-PacF4	8	Yes	1
6924	SE	EI	TSL	AJI_1	1	No	1
7417	SE	EI	TSL	AJI_7+OF-PacF7	8	Yes	1
9536	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
7063	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9543	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9545	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9549	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9583	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9837	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9871	SE	UNC	UNC-HTSF	NEBNext_E7335	2	No	1
9944	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
6974	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	1
9555	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	NA
9598	SE	UNC	UNC-HTSF	NEBNext_E7335	1	No	NA
9905	SE	UNC	UNC-HTSF	NEBNext_E7335	2	No	NA
9762	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9764	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9332	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9869	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	4	No	1
9879	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9887	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9947	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	2	No	1
9832	SE	UNC	UNC-HTSF/MOgene	NEBNext_E7335	4	No	NA

Table A.3.: Survey of NLR gene numbers for 92 species reported in 14 different papers. Species names, NLR gene numbers, and reported NLR subclasses for all reference papers are given.

Species	NLRs	NLR subclasses reported	Reference
Actinidia chinensis	90	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Amborella trichopoda	105	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Amborella trichopoda	88	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Ananas comosus	164	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Aquilegia caerulea	195	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Arabidopsis halleri	135	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Arabidopsis lyrata	241	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Arabidopsis lyrata	134	TNL,CNL,TN	(Peele et al. 2014)
Arabidopsis lyrata	204	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Arabidopsis lyrata	198	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Arabidopsis lyrata	202	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Arabidopsis lyrata	185	CNL,CN,TNL,TN,NL,N	(Guo et al. 2011)
Arabidopsis thaliana	238	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Arabidopsis thaliana	135	TNL,CNL,TN	(Peele et al. 2014)

Table A.3.: continued

Species	NLRs	NLR subclasses reported	Reference
Arabidopsis thaliana	165	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Arabidopsis thaliana	213	TNL,CNL,NL,TN,CN,N	(Yu et al. 2014)
Arabidopsis thaliana	168	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Arabidopsis thaliana	213	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Arabidopsis thaliana	149	CNL,NL,TNL,TN,CN,C,N,Others	(Meyers et al. 2003)
Arabidopsis thaliana	149	CNL,CN,TNL,TN,NL,N	(Guo et al. 2011)
Asparagus officinalis	54	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Beta vulgaris	117	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Boechera strica	314	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Brachypodium distachyon	185	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Brachypodium distachyon	501	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Brachypodium distachyon	253	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Brachypodium distachyon	327	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Brassica napus	499	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Brassica napus	471	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Brassica oleracea	239	TNL,CNL,NL,TN,CN,N	(Yu et al. 2014)
Brassica oleracea	327	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Brassica rapa	204	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Brassica rapa	248	TNL,CNL,NL,TN,CN,N	(Yu et al. 2014)
Brassica rapa	151	TNL,CNL,TN	(Peele et al. 2014)
Brassica rapa	207	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Brassica rapa	196	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Cajanus cajan	256	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Cajanus cajan	289	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016, 2014)
Cajanus cajan	815	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Capsella grandiflora	115	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Capsella grandiflora	103	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Capsella rubella	75	TNL,CNL,TN	(Peele et al. 2014)
Capsella rubella	152	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Capsella rubella	127	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Capsella rubella	131	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Capsicum annuum	305	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Capsicum annuum	661	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Carica papaya	48	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Carica papaya	56	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Carica papaya	46	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Chlamydomonas reinhardtii	0	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Chlamydomonas reinhardtii	0	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Cicer arietinum	102	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Cicer arietinum	227	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Citrullus lanatus	42	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Citrus clementina	399	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Citrus sinensis	443	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Coccomyxa subellipsoidea	0	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Coccomyxa subellipsoidea	3	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Coffea canephora	715	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Cucumis sativus	79	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Cucumis sativus	76	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Cucumis sativus	65	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Elaeis guineensis	145	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Eucalyptus grandis	872	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Eucalyptus grandis	788	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Eutrema salsugineum	67	TNL,CNL,TN	(Peele et al. 2014)
Eutrema salsugineum	136	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Eutrema salsugineum	125	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Fragaria vesca	190	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Fragaria vesca	164	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Fragaria vesca	144	TNL,n-TNL,CNL,XNL	(Zhong et al. 2015)
Fraxinus excelsior	173	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Glycine max	325	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Glycine max	784	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Glycine max	465	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016, 2014)
Glycine max	442	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Glycine max	744	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Glycine soja	952	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Gossypium raimondii	369	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Gossypium raimondii	289	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Hordeum vulgare	336	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Hordeum vulgare	318	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Jatropha curcas Palawan	185	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Leavenworthia alabamica	121	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Linum usitatissimum	196	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Linum usitatissimum	168	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Lotus japonicus	176	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Lotus japonicus	247	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Lotus japonicus	270	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Malus domestica	1032	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Malus domestica	935	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Malus domestica	748	TNL,n-TNL,CNL,XNL	(Zhong et al. 2015)
Manihot esculenta	148	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Manihot esculenta	232	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Manihot esculenta	212	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Medicago truncatula	668	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Medicago truncatula	1074	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Medicago truncatula	753	TNL,CNL,NL,TN,CN,N	(Yu et al. 2014)
Medicago truncatula	571	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016, 2014)

Table A.3.: continued

Species	NLRs	NLR subclasses reported	Reference
Medicago truncatula	771	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Medicago truncatula	770	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Micromonas pusilla	0	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Micromonas pusilla	0	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Micromonas pusilla	0	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Mimulus guttatus	138	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Mimulus guttatus	344	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Mimulus guttatus	317	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Musa acuminata	111	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Musa acuminata	105	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Musa balbisiana	85	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Nelumbo nucifera	115	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Nicotiana tabacum	952	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Oryza glaberrima	292	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Oryza sativa	595	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Oryza sativa	498	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Oryza sativa	470	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Oryza sativa indica	616	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Oryza sativa japonica	578	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Ostreococcus lucimarinus	0	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Ostreococcus lucimarinus	0	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Panicum virgatum	1005	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Phalaenopsis equestris	60	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Phaseolus vulgaris	406	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Phaseolus vulgaris	337	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016, 2014)
Phaseolus vulgaris	334	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Phaseolus vulgaris	359	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Phyllostachys edulis	311	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Phyllostachys heterocycla	344	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Physcomitrella patens	44	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Physcomitrella patens	87	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Physcomitrella patens	69	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Picea abies	562	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Populus trichocarpa	369	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Populus trichocarpa	728	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Populus trichocarpa	483	TNL,CNL,NL,TN,CN,N	(Yu et al. 2014)
Populus trichocarpa	554	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Prunus mume	352	TNL,n-TNL,CNL,XNL	(Zhong et al. 2015)
Prunus persica	435	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Prunus persica	416	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Prunus persica	399	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Prunus persica	354	TNL,n-TNL,CNL,XNL	(Zhong et al. 2015)
Pyrus bretschneideri	469	TNL,n-TNL,CNL,XNL	(Zhong et al. 2015)
Ricinus communis	169	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Ricinus communis	167	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Ricinus communis	139	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Selaginella moellendorffii	12	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Selaginella moellendorffii	18	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Selaginella moellendorffii	16	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Sesamum indicum	170	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Setaria italica	470	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Setaria italica	424	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Setaria italica	380	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Sisymbrium irio	218	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Solanum lycopersicum	264	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Solanum lycopersicum	255	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Solanum lycopersicum	223	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Solanum lycopersicum	355	CNL,TNL,CN,TL,TN,N,T,L	(Andolfo et al. 2014)
Solanum melongena	242	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Solanum pennellii	201	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Solanum tuberosum	543	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Solanum tuberosum	447	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Solanum tuberosum	355	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Solanum tuberosum	438	TN,TNL,CN,CNL,NL,N	(Jupe et al. 2012)
Sorghum bicolor	317	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Sorghum bicolor	422	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Sorghum bicolor	326	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Sorghum bicolor	310	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Spinacia oleracea	58	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Thellungiella halophila	131	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Thellungiella salsuginea	88	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Theobroma cacao	355	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Theobroma cacao	314	TNL,CNL,NL,TN,CN,N	(Yu et al. 2014)
Theobroma cacao	273	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Triticum aestivum	1224	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Triticum aestivum	1077	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Triticum urartu	563	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Vitis vinifera	590	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)
Vitis vinifera	323	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Vitis vinifera	545	TNL,CNL,NL,TN,CN,N	(Yu et al. 2014)
Vitis vinifera	295	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Vitis vinifera	314	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Vitis vinifera	754	TNL,TN,T,CNL,CN,C,CTN,L,NL,N,Others	(Zheng et al. 2016)
Volvox carteri	2	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Volvox carteri	0	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)
Zea mays	122	CNL,CN,Nlcc,Ncc,TNL,TN,NLTIR,Tx	(Kim et al. 2012)

Table A.3.: continued

Species	NLRs	NLR subclasses reported	Reference
Zea mays	191	NLR,N,TLR,TN,NB-ID,T2N	(Sarris et al. 2016)
Zea mays	139	TNL,TN,NL,N,CNL,CN,RNL,RN	(Shao et al. 2016)
Zea mays	129	TNL,TN,CNL,CN,RNL,RN,NL,N	(Zhang et al. 2016)

Table A.4.: NLRs showing at least 1.5x increased expected mean normalized coverage.

norm. coverage	NLR	architecture	subclass	OG	OG size	Col-0 NLR in OG	assembly Qual.
6.3796148475	7396|T234-R1	RPW8	R	2.26	3	-	30.5467
5.7928488541	7396|T235-R1	RPW8	R	108.1	13	-	30.5467
4.7575541689	9542|T392-R1	TIR,NB,LRR	TNL	180.5	2	-	33.4558
4.5255839096	7396|T236-R1	RPW8	R	2.9	24	-	30.5467
3.7456716058	9944|T166-R1	TIR,NB,LRR	TNL	5.1	4	-	35.1517
3.6816898296	7396|T233-R1	RPW8	R	151.1	11	-	30.5467
3.6724871713	7322|T388-R1	RPW8	R	2.21	12	-	32.5242
3.3421473072	9869|T322-R1	TIR,NB,LRR	TNL	165.5	15	-	35.4059
3.1510536671	9944|T414-R1	RPW8	R	2.12	3	-	35.1517
3.0994992479	9610|T512-R1	NB,LRR	NL	842	1	-	37.6482
3.0842557324	7396|T232-R1	RPW8	R	2.21	12	-	30.5467
3.0583708295	9658|T314-R1	Coil,NB	CN	197.1	23	-	36.168
3.0233098354	9557|T421-R1	RPW8	R	151.1	10	-	37.2282
2.9740426995	108|T522-R1	TIR,NB	TN	239.1	14	-	35.3142
2.9715010293	108|T375-R1	RPW8	R	108.16	9	-	35.3142
2.9514531872	9837|T471-R1	NB,Coil	CN	290.1	5	-	33.4813
2.9463047237	7308|T133-R1	RPW8	R	151.1	10	-	26.3469
2.9350211606	9554|T447-R1	RPW8	R	108.12	9	-	34.59
2.9042692299	6924|T201-R1	TIR,NB	TN	1418	1	-	31.8952
2.8598006773	7322|T240-R1	RPW8	R	108.1	13	-	32.5242
2.8208349621	9658|T315-R1	NB	N	208.1	21	-	36.168
2.8144623532	7322|T387-R1	RPW8	R	151.1	11	-	32.5242
2.7907815413	9550|T042-R1	TIR	T	178.6	2	-	33.5993
2.6834291108	9332|T245-R1	RPW8	R	2.9	24	-	26.5922
2.6573671726	6924|T121-R1	NB	N	324.1	2	-	31.8952
2.5584151135	9536|T151-R1	RPW8	R	151.1	10	-	34.7751
2.4995018919	9597|T258-R1	Coil,NB	CN	15.6	6	-	35.0453
2.4752911123	7328|T223.2-R1	TIR,NB,LRR	TNL	16.6	12	-	29.8617
2.4745144008	5993|T426-R1	TIR	T	11.18	3	-	34.2887
2.4201812456	9721|T186-R1	RPW8	R	108.12	9	-	35.6259
2.4119991367	9536|T152-R1	RPW8	R	2.31	9	-	34.7751
2.4115434378	9600|T253-R1	RPW8	R	108.12	9	-	34.3225
2.4067866467	7288|T309-R1	TIR,NB,LRR	TNL	258.1	9	-	27.4731
2.3856027279	7067|T356-R1	TIR	T	248.1	11	-	30.9625
2.3773175962	9542|T372-R1	TIR	T	45.5	15	-	33.4558
2.3690469027	7328|T231-R1	RPW8	R	2.3	10	-	29.8617
2.3169125918	7328|T701-R1	TIR	T	11.14	4	-	29.8617
2.2978494015	9518|T536-R1	TIR	T	11.14	4	-	39.9199
2.2694942564	7396|T427-R1	Coil,NB	CN	1039	1	-	30.5467
2.2446125075	9554|T031.1-R1	TIR,NB,LRR	TNL	16.11	10	-	34.59
2.1788578288	9944|T413-R1	RPW8	R	108.1	13	-	35.1517
2.1607044513	108|T370-R1	Coil,NB	CN	15.1	8	-	35.3142
2.1514656919	6906|T345-R1	RPW8	R	108.14	3	-	34.1607
2.110864828	9762|T004-R1	NB	N	234.1	11	-	34.6609
2.0992539035	9869|T028-R1	TIR,NB,LRR	TNL	241.1	14	AT4G16960.1	35.4059
2.0717831401	9871|T078-R1	TIR,NB,LRR	TNL	176.5	2	-	36.2821
2.0657780394	9554|T340-R1	RPW8	R	2.17	8	-	34.59
2.0393804468	9762|T005-R1	LRR,TIR	TL	178.2	2	-	34.6609
2.00983519	9550|T273-R1	TIR,NB,LRR	TNL	241.1	14	AT4G16960.1	33.5993
2.0039150215	7308|T134-R1	RPW8	R	265.1	8	-	26.3469
2.0032320573	9550|T553-R1	TIR,NB,LRR	TNL	180.2	16	-	33.5993
1.9984046747	7058|T252-R1	RPW8	R	2.1	7	-	31.6562
1.9924461234	9100|T098-R1	TIR,NB,LRR	TNL	4.8	19	-	35.6972
1.9784204106	9654|T448-R1	TIR,NB,LRR	TNL	180.2	16	-	36.6904
1.9771357215	9762|T011-R1	TIR	T	799	3	-	34.6609
1.9767286352	7328|T389-R1	Coil,NB,LRR	CNL	97.1	56	AT1G51480.1	29.8617
1.9621041881	9871|T537-R1	TIR	T	11.16	7	-	36.2821
1.9509814519	9911|T545-R1	TIR	T	11.6	2	-	37.3617
1.9488038407	7396|T417-R1	NB,LRR	NL	1220	1	-	30.5467
1.9374840712	7415|T062-R1	Coil,NB,LRR	CNL	7.21	1	-	31.3233
1.935526326	7111|T233-R1	RPW8	R	2.1	1	-	30.1482
1.9329574281	6981|T548-R1	TIR	T	11.23	4	-	38.3333
1.9292594974	9869|T026-R1	TIR,NB,LRR	TNL	178.4	4	-	35.4059
1.9258505424	9762|T008-R1	TIR,NB,LRR	TNL	4.16	19	AT4G16920.2	34.6609
1.9214469491	9134|T514-R1	TIR	T	11.2	9	-	33.1895
1.9033161959	9837|T552-R1	TIR	T	11.16	7	-	33.4813
1.9007978579	9792|T497-R1	TIR	T	11.19	2	-	38.7705
1.8934859275	9721|T634-R1	TIR	T	11.2	9	-	35.6259
1.8892769722	6909|T487-R1	TIR	T	-	-	-	42.8018
1.8885708741	9537|T478-R1	TIR	T	11.2	9	-	33.7539
1.8514805533	6899|T539-R1	TIR	T	11.1	2	-	31.8748
1.8415880027	9669|T187-R1	TIR,NB,LRR	TNL	180.2	16	-	36.6465
1.8340551546	6974|T551-R1	TIR	T	11.6	2	-	33.7577
1.8300150201	5993|T089-R1	TIR,NB,LRR	TNL	163.7	5	-	34.2887
1.8230143621	7288|T332-R1	RPW8	R	108.16	9	-	27.4731

Table A.4.: continued

norm. cov.	NLR	architecture	subclass	OG	OG size	Col-0 NLR in OG	assembly Qual.
1.8224571158	9762|T001-R1	TIR,NB,LRR	TNL	1697	1	-	34.6609
1.8220499435	9580|T696-R1	TIR	T	11.2	4	-	33.4954
1.8204749852	9762|T082-R1	TIR	T	766	2	-	34.6609
1.8175287288	7167|T322-R1	TIR,NB,LRR, PF01451	TNLX	180.2	16	-	30.6264
1.8076622406	108|T560-R1	TIR	T	11.25	2	-	35.3142
1.797834481	9669|T186-R1	TIR	T	248.1	11	-	36.6465
1.7920624571	9557|T422-R1	RPW8	R	2.31	9	-	37.2282
1.7907483271	7373|T324-R1	TIR	T	11.2	9	-	31.182
1.7899058415	9762|T002-R1	TIR	T	266.1	9	-	34.6609
1.7872082951	9871|T288-R1	RPW8	R	108.8	6	-	36.2821
1.7788454329	9879|T128.1-R1	TIR	T	118.1	50	AT5G38344.1	33.5256
1.7687549604	7058|T112-R1	Coil,NB,LRR	CNL	3.17	7	-	31.6562
1.7685367737	9597|T566-R1	TIR	T	11.2	9	-	35.0453
1.7589981208	9332|T332-R1	NB	N	324.1	2	-	26.5922
1.7513540671	9837|T224-R1	RPW8	R	151.8	2	-	33.4813
1.7390433737	9871|T474-R1	TIR	T	98.1	56	AT1G47370.1 (RBA1)	36.2821
1.7377237426	9944|T451-R1	Coil,NB,LRR	CNL	7.15	2	-	35.1517
1.7319366831	7063|T464-R1	TIR	T	11.22	2	-	35.8094
1.7297532329	6939|T320-R1	TIR,NB,LRR	TNL	180.4	3	-	27.9431
1.7201943502	9762|T003-R1	TIR	T	283.1	7	-	34.6609
1.713378543	9764|T236-R1	NB	N	12.7	10	-	37.0734
1.7130822354	7396|T401-R1	TIR	T	11.2	9	-	30.5467
1.708403574	7273|T077-R1	TIR,NB	TN	122.1	54	AT1G72900.1	31.7571
1.7068195328	6909|T337-R1	NB	N	-	-	-	42.8018
1.7038458308	9610|T630-R1	TIR	T	11.2	9	-	37.6482
1.6794811972	9600|T549-R1	TIR,LRR	TL	1157	1	-	34.3225
1.6788728093	6924|T200-R1	TIR,NB	TN	24.1	63	AT1G27170.1	31.8952
1.6684371159	9871|T289-R1	RPW8	R	2.9	24	-	36.2821
1.6651488698	9837|T051-R1	PF03106,NB,LRR, PF00069	NLX	41.1	59	AT4G12020.2 (MAPKKK11)	33.4813
1.6562499455	7111|T044-R1	NB,LRR	NL	788	2	-	30.1482
1.6462368984	7288|T093-R1	Coil,NB,LRR	CNL	7.8	6	-	27.4731
1.6443940279	6924|T457-R1	TIR	T	11.13	2	-	31.8952
1.6413884627	9837|T385-R1	TIR,NB,LRR	TNL	10.8	16	-	33.4813
1.6333724979	6974|T290-R1	RPW8	R	108.1	13	-	33.7577
1.6120713487	9543|T532-R1	TIR	T	11.16	7	-	33.3815
1.6075536	7308|T339-R1	RPW8	R	108.16	9	-	26.3469
1.6053856669	7328|T729-R1	TIR	T	1562	1	-	29.8617
1.5945431782	7213|T269-R1	RPW8	R	2.21	12	-	29.9984
1.5942648897	5784|T006.2-R1	PF02671,PF03106, TIR,NB,LRR	TNLX	41.1	59	AT4G12020.2 (MAPKKK11)	34.479
1.5921338902	7111|T234-R1	RPW8	R	108.1	13	-	30.1482
1.5758152716	7111|T043-R1	TIR,NB,LRR	TNL	1286	1	-	30.1482
1.5650071293	5784|T183-R1	Coil,NB	CN	8.1	1	-	34.479
1.5502151979	9597|T034.1-R1	PF02671,PF03106, TIR,NB,LRR	TNLX	41.1	59	AT4G12020.2 (MAPKKK11)	35.0453
1.548812712	9597|T334-R1	NB,LRR	NL	873	1	-	35.0453
1.5466139936	9518|T002-R1	TIR,NB,LRR	TNL	1.27	26	-	39.9199
1.5438567172	6909|T327-R1	NB	N	-	-	-	42.8018
1.5433065112	7067|T247.2-R1	PF02671,PF03106, TIR,NB,LRR	TNLX	41.1	59	AT4G12020.2 (MAPKKK11)	30.9625
1.5403231487	5993|T357-R1	TIR,NB,LRR	TNL	206.1	20	-	34.2887
1.5394101809	9792|T166-R1	RPW8	R	108.12	9	-	38.7705
1.5337312161	9869|T213-R1	TIR,NB,LRR	TNL	221.1	2	-	35.4059
1.5303968244	7186|T252-R1	RPW8	R	2.13	6	-	29.3276
1.5285793405	108|T571-R1	TIR	T	192.1	27	-	35.3142
1.5258943958	9654|T475-R1	RPW8	R	2.21	12	-	36.6904
1.5221585901	9518|T338-R1	TIR,NB,LRR	TNL	165.5	15	-	39.9199
1.5212131775	5784|T410.2-R1	TIR,NB	TN	164.1	37	AT1G27180.1	34.479
1.5169843923	7067|T085.2-R1	TIR,NB,LRR	TNL	1.14	11	-	30.9625
1.515837007	9600|T436-R1	TIR,NB,LRR	TNL	5.2	9	AT5G41740.2 (SSI4-like)	34.3225
1.5110189653	1925|T339-R1	TIR,NB	TN	239.1	14	-	37.8127
1.5099428468	7322|T199-R1	TIR,NB,LRR	TNL	5.1	26	-	32.5242
1.5079159765	7328|T221-R1	TIR,NB,LRR	TNL	5.12	19	-	29.8617
1.5075480622	9669|T534-R1	TIR	T	11.15	1	-	36.6465
1.501615648	7288|T207-R1	TIR,NB,LRR	TNL	176.1	2	-	27.4731